

Jacques Blanc-Talon
Wilfried Philips
Dan Popescu
Paul Scheunders (Eds.)

LNCS 4179

Advanced Concepts for Intelligent Vision Systems

8th International Conference, ACIVS 2006
Antwerp, Belgium, September 2006
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

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

Moshe Y. Vardi

Rice University, Houston, TX, USA

Gerhard Weikum

Max-Planck Institute of Computer Science, Saarbruecken, Germany

Jacques Blanc-Talon Wilfried Philips
Dan Popescu Paul Scheunders (Eds.)

Advanced Concepts for Intelligent Vision Systems

8th International Conference, ACIVS 2006
Antwerp, Belgium, September 18-21, 2006
Proceedings

Volume Editors

Jacques Blanc-Talon
DGA/D4S/MRIS
16 bis, avenue Prieur de la Côte d'Or, 94114 Arcueil, France
E-mail: jacques.blanc-talon@etca.fr

Wilfried Philips
Ghent University
Telecommunication and Information Processing
St.-Pietersnieuwstraat 41, 9000 Ghent, Belgium
E-mail: philips@ugent.be

Dan Popescu
CSIRO ICT Centre
P.O. Box 76, Epping, NSW 1710, 1710 Sydney, Australia
E-mail: Dan.Popescu@csiro.au

Paul Scheunders
University of Antwerp
Building N. 2610 Wilrijk, Universiteitsplein 1, 2610 Wilrijk, Belgium
E-mail: paul.scheunders@ua.ac.be

Library of Congress Control Number: 2006932485

CR Subject Classification (1998): I.4, I.5, I.3, I.2, I.2.10

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

ISSN 0302-9743
ISBN-10 3-540-44630-3 Springer Berlin Heidelberg New York
ISBN-13 978-3-540-44630-9 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 2006
Printed in Germany

Typesetting: Camera-ready by author, data conversion by Scientific Publishing Services, Chennai, India
Printed on acid-free paper SPIN: 11864349 06/3142 5 4 3 2 1 0

Preface

This volume collects the papers accepted for presentation at the Eight International Conference on “Advanced Concepts for Intelligent Vision Systems” (Acivs 2006). The ACIVS conference was established in 1999 in Baden-Baden (Germany) as part of a large multiconference. ACIVS has maintained the tradition of being a single track event with oral presentations 25 minutes each, even though the number of participants has been steadily growing every year. The conference currently attracts computer scientists from more than 20 countries, mostly from Europe, Australia and Japan, but also from the USA, Asia and the Middle-East.

Though ACIVS is a conference on all areas of image processing, one of its major domains is image and video compression. A third of the selected papers dealt with compression, motion estimation, moving object detection and other video applications. This year, topics related to clustering, pattern recognition and biometrics constituted another third of the conference. The last third was more related to the fundamentals of image processing, namely noise reduction, filtering, restoration and image segmentation. We would like to thank the invited speakers Kathrin Berkner (Ricoh Innovations), Nikos Paragios (Ecole Centrale de Paris) and Dimitri Van de Ville (Ecole Polytechnique Federale de Lausanne) for enhancing the technical program with their presentations.

A conference like ACIVS would not be feasible without the concerted effort of many people and support of various institutions. The paper submission and review procedure was carried out electronically and a minimum of 3 reviewers were assigned to every paper. From 242 submissions, 45 were selected for oral presentation and 81 as posters. A large and energetic Program Committee, helped by additionnal referees – listed on the following pages – completed the long and demanding reviewing process. We would like to thank all of them for their timely and high-quality reviews. Also, we would like to thank our sponsors, Philips Research, Barco, Eurasip, the IEEE Benelux Signal Processing Chapter and the Flemish FWO Research Community on Image Processing Systems, for their valuable support.

Last but not least, we would like to thank all the participants who trusted us in organizing this event for the seventh time. We hope they attended a stimulating scientific event and enjoyed the atmosphere of the ACIVS social events in the historic city of Antwerp.

Organization

Acivs 2006 was organized by the University of Antwerp and Ghent University.

Steering Committee

Jacques Blanc-Talon (DGA/D4S/MRIS, Arcueil, France)
Wilfried Philips (Ghent University, Ghent, Belgium)
Dan Popescu (CSIRO, Sydney, Australia)
Paul Scheunders (University of Antwerp, Wilrijk, Belgium)

Organizing Committee

Wilfried Philips (Ghent University, Ghent, Belgium)
Paul Scheunders (University of Antwerp, Wilrijk, Belgium)

Sponsors

Acivs 2006 was sponsored by the following organizations:

- Faculty of Engineering Sciences, Ghent University
- Philips Research
- The IEEE Benelux Signal Processing Chapter
- Eurasip
- Barco
- DSP Valley
- The FWO Research Community on Audiovisual Systems (AVS)

The Acivs 2006 organizers are especially grateful to Philips Research for their financial sponsorship. They are also grateful to the FWO Research Community on Audiovisual Systems for sponsoring some of the invited speakers and to Barco for providing a small present for the participants.

Program Committee

Fritz Albregtsen (University of Oslo, Oslo, Norway)
Attila Baskurt (INSA Lyon, Villeurbanne, France)
Laure Blanc-Feraud (CNRS, Sophia-Antipolis, France)
Philippe Bolon (University of Savoie, Annecy, France)
Nikolaos Bourbakis (Wright State University, Dayton, USA)

Salah Bourennane (EGIM, Marseille, France)
Patrick Bouthemy (IRISA/INRIA, Rennes, France)
Jocelyn Chanussot (INPG, Grenoble, France)
David Clausi (University of Waterloo, Waterloo, Canada)
Pamela Cosman (University of California at San Diego, La Jolla, USA)
Jennifer Davidson (Iowa State University, Ames, USA)
Ricardo de Queiroz (Universidade de Brasilia, Brasilia, Brazil)
Christine Fernandez-Maloigne (Université de Poitiers, Chasseneuil, France)
Jan Flusser (Institute of Information Theory and Automation, Prague, Czech Republic)
Don Fraser (University of New South Wales, Canberra, Australia)
Sidharta Gautama (Ghent University, Ghent, Belgium)
Jérôme Gilles (CEP, Arcueil, France)
Georgy Gimel'farb (The University of Auckland, Auckland, New Zealand)
Daniele Giusto (University of Cagliari, Cagliari, Italy)
Mark Huiskes (CWI, Amsterdam, the Netherlands)
John Illingworth (University of Surrey, Guildford, UK)
Pieter Jonker (Delft University of Technology, the Netherlands)
Frédéric Jurie (CNRS - INRIA, Saint Ismier, France)
Andrzej Kasinski (Poznan University of Technology, Poznan, Poland)
Ashraf Kassim (National University of Singapore, Singapore, Singapore)
Richard Kleihorst (Philips Research, Eindhoven, the Netherlands)
Murat Kunt (EPFL, Lausanne, Switzerland)
Hideo Kuroda (Nagasaki University, Nagasaki, Japan)
Kenneth Lam (The Hong Kong Polytechnic University, Hong Kong, China)
Bruce Litow (James Cook University, Townsville, Australia)
Rastislav Lukac (University of Toronto, Toronto, Canada)
Jesus Malo (Universitat de Valencia, Burjassot, Spain)
Gérard Medioni (USC/IRIS, Los Angeles, USA)
Fabrice Mériaudeau (IUT Le Creusot, Le Creusot, France)
Ali Mohammad-Djafari (CNRS, Gif-sur-Yvette, France)
Rafael Molina (Universidad de Granada, Granada, Spain)
Vittorio Murino (Università degli Studi di Verona, Verona, Italy)
Edgard Nyssen (Vrije Universiteit Brussel, Brussels, Belgium)
Stanley Osher (UCLA, Los Angeles, USA)
Marcin Paprzycki (SWPS, Warsaw, Poland)
Jussi Parkkinen (University of Joensuu, Joensuu, Finland)
Fernando Pereira (Instituto Superior Técnico, Lisboa, Portugal)
Béatrice Pesquet-Popescu (ENST, Paris, France)
Matti Pietikäinen (University of Oulu, Oulu, Finland)
Aleksandra Pizurica (Ghent University, Ghent, Belgium)
Gianni Ramponi (Trieste University, Trieste, Italy)
Alpesh Kumar Ranchordas (FMx Ltd, Crawley, UK)

Paolo Remagnino (Faculty of Technology, Kingston University, Surrey, UK)
 Luis Salgado Álvarez de Sotomayor (Universidad Politécnica de Madrid,
 Madrid, Spain)
 Frederic Truchetet (Université de Bourgogne, Le Creusot, France)
 Dimitri Van De Ville (EPFL, Lausanne, Switzerland)
 Peter Veelaert (University College Ghent, Ghent, Belgium)

Reviewers

Arnaldo Abrantes (ISEL, Lisbon, Portugal)
 Fritz Albreghsen (University of Oslo, Oslo, Norway)
 Jesus Angulo (Ecole des Mines de Paris, Fontainebleau, France)
 Didier Auroux (Université Paul Sabatier, Toulouse, France)
 Raphaèle Balter (France Télécom, Rennes, France)
 Attila Baskurt (INSA Lyon, Villeurbanne, France)
 Azzedine Beghdadi (Institut Galilée, Villetaneuse, France)
 Abdel Belaid (LORIA, Vandoeuvre les Nancy, France)
 Angelo Beraldin (NRC Institute for Information Technology, Ottawa, Canada)
 Thierry Berger (Université de Limoges, Limoges, France)
 Kathrin Berkner (Ricoh Innovations, Menlo Park, USA)
 Gilles Bertrand (ESIEE, Marne-la-Vallée, France)
 Jens Bialkowski (Universität Erlangen-Nürnberg, Erlangen, Germany)
 Laure Blanc-Feraud (CNRS, Sophia-Antipolis, France)
 Jacques Blanc-Talon (DGA/D4S/MRIS, Arcueil, France)
 Isabelle Bloch (Ecole Nationale Supérieure des Telecommunications,
 Paris, France)
 Leonardo Bocchi (University of Florence, Florence, Italy)
 Philippe Bolon (University of Savoie, Annecy, France)
 Patrick Bonnin (Université de Versailles, Velizy, France)
 Alberto Borghese (University of Milan, Milan, Italy)
 Faysal Boughorbel (Philips Research, Eindhoven, the Netherlands)
 Nikolaos Bourbakis (Wright State University, Dayton, USA)
 Salah Bourennane (EGIM, Marseille, France)
 Pierrick Bourgeat (CSIRO, Sydney, Australia)
 Patrick Bouthemy (IRISA/INRIA, Rennes, France)
 Frederic Champagnat (Office National D'Etudes et Recherches Aerospatiales,
 Chatillon, France)
 Alex Chan (US Army Research Laboratory, Adelphi, USA)
 Jocelyn Chanussot (INPG, Grenoble, France)
 Jean-Marc Chassery (INPG, Grenoble, France)
 Sei-Wang Chen (National Taiwan Normal University, Taipei, Taiwan)
 Shu-Yuan Chen (Yuan Ze University, Taoyuan, Taiwan)
 Xilin Chen (Chinese Academy of Sciences, Beijing, China)
 Sen-ching Cheung (University of Kentucky, Lexington, USA)
 Emmanuel Christophe (CNES, Toulouse, France)

Petr Cimprich (Ginger Alliance, Prague, Czech Republic)
David Clausi (University of Waterloo, Waterloo, Canada)
Mary Comer (Purdue University, West Lafayette, USA)
Pamela Cosman (University of California at San Diego, La Jolla, USA)
Wim d' Haes (University of Antwerp, Antwerp, Belgium)
Emiliano D'Agostino (KULeuven, Leuven, Belgium)
Arthur da Cunha (University of Illinois at Urbana-Champaign,
Urbana-Champaign, USA)
Matthew Dailey (Asian Institute of Technology, Klong Luang, Thailand)
André Dalgarrondo (DGA / CEV, Cazaux, France)
Frederic Dambreville (CEP, Arcueil, France)
Jennifer Davidson (Iowa State University, Ames, USA)
Steve De Backer (University of Antwerp, Wilrijk, Belgium)
Johan De Bock (Ghent University, Ghent, Belgium)
Arturo de la Escalera (Universidad Carlos III de Madrid, Leganes, Spain)
Ricardo de Queiroz (Universidade de Brasilia, Brasilia, Brazil)
Patrick De Smet (Ghent University, Ghent, Belgium)
Arturo del Bimbo (Università degli Studi di Firenze, Firenze)
Stéphane Derrode (EGIM, Marseille, France)
Agnès Desolneux (CNRS, Université Paris 5, Paris, France)
Bart Dhoedt (Ghent University, Ghent, Belgium)
Jef Driesen (University of Antwerp, Wilrijk, Belgium)
Hans Du Buf (University of Algarve, Faro, Portugal)
Delphine Dufourd (DGA, Paris, France)
Dirk Farin (TU-Eindhoven, Eindhoven, the Netherlands)
Hamed Fatemi (Eindhoven University, Eindhoven, the Netherlands)
Christine Fernandez-Maloigne (Université de Poitiers, Chasseneuil, France)
Jan Flusser (Institute of Information Theory and Automation, Prague,
Czech Republic)
Don Fraser (University of New South Wales, Canberra, Australia)
Denis Friboulet (Creatis, Villeurbanne, France)
André Gagalowicz (INRIA, Rocquencourt, France)
Christoph Garbe (IWR, University of Heidelberg, Heidelberg, Germany)
Sidharta Gautama (Ghent University, Ghent, Belgium)
Jérôme Gilles (CEP, Arcueil, France)
Georgy Gimel'farb (The University of Auckland, Auckland, New Zealand)
Daniele Giusto (University of Cagliari, Cagliari, Italy)
Hervé Glotin (Lab. System & Information Sciences-UMR CNRS, La Garde,
France)
Bart Goethals (University of Antwerp, Antwerp, Belgium)
Philippe-Henri Gosselin (ENSEA, Cergy-Pontoise, France)
Valérie Gouet-Brunet (Conservatoire National des Arts et Métiers, Paris, France)
D.S. Guru (University of Mysore, Mysore, India)
Rudolf Hanel (University of Antwerp, Antwerp, Belgium)

Vaclav Hlavac (Czech Technical University, Faculty of Electrical Engineering, Prague 2, Czech Republic)

Mark Holden (CSIRO ICT Centre, Sydney, Australia)

Mark Huiskes (CWI, Amsterdam, the Netherlands)

Bruno Huysmans (Ghent University, Ghent, Belgium)

Osamu Ikeda (Takushoku University, Tokyo, Japan)

John Illingworth (University of Surrey, Guildford, UK)

Maarten Jansen (TU Eindhoven, Eindhoven, the Netherlands)

Philippe Joly (Université Paul Sabatier - IRIT, Toulouse, France)

Pieter Jonker (Delft University of Technology, the Netherlands)

Frédéric Jurie (CNRS - INRIA, Saint Ismier, France)

Stavros Karkanis (Technological Educational Institute (TEI) of Lamia, Lamia, Greece)

Andrzej Kasinski (Poznan University of Technology, Poznan, Poland)

Ashraf Kassim (National University of Singapore, Singapore, Singapore)

Ali Khenchaf (ENSIETA, Brest, France)

Cheong Ghil Kim (Yonsei University, Seoul, Korea)

Joseph Kittler (University of Surrey, Cambridge, UK)

Richard Kleihorst (Philips Research, Eindhoven, the Netherlands)

Stephan Kopf (Mannheim University, Mannheim, Germany)

Cris Koutsougeras (Tulane University, New Orleans, USA)

Murat Kunt (EPFL, Lausanne, Switzerland)

Hideo Kuroda (Nagasaki University, Nagasaki, Japan)

Vivek Kwatra (University of North Carolina at Chapel Hill, Chapel Hill, USA)

Olivier Laligant (Le2i Lab., Le Creusot, France)

Kenneth Lam (The Hong Kong Polytechnic University, Hong Kong, China)

Patrick Lambert (ESIA, Annecy, France)

Peter Lambert (Ghent University, Ledeborg-Ghent, Belgium)

Ivan Laptev (INRIA, Rennes, France)

Alessandro Ledda (Ghent University, Ghent, Belgium)

Alexander Leemans (University of Antwerp, Wilrijk, Belgium)

Sébastien Lefèvre (University Louis Pasteur - Strasbourg 1, Illkirch, France)

Rongxin Li (CSIRO ICT Centre, Epping, NSW, Australia)

Chia-Wen Lin (National Chung Cheng University, Chiayi, Taiwan)

Bruce Litow (James Cook University, Townsville, Australia)

Rastislav Lukac (University of Toronto, Toronto, Canada)

Hiep Luong (Ghent University, Ghent, Belgium)

Evelyne Lutton (INRIA, Le Chesnay, France)

Dominique Luzeaux (DGA, Arcueil, France)

Antoine Manzanera (ENSTA, Paris, France)

Kirk Martinez (The University of Southampton, Southampton, UK)

David Masip (Computer Vision Center, Bellaterra, Spain)

Gérard Medioni (USC/IRIS, Los Angeles, USA)

Fabrice Mériaudeau (IUT Le Creusot, Le Creusot, France)

Maurice Milgram (Jussieu Université, Paris, France)

Ali Mohammad-Djafari (CNRS, Gif-sur-Yvette, France)
Rafael Molina (Universidad de Granada, Granada, Spain)
Greg Mori (Simon Fraser University, Burnaby, Canada)
Adrian Munteanu (Vrije Universiteit Brussel, Brussels, Belgium)
Vittorio Murino (Università degli Studi di Verona, Verona, Italy)
Mike Nachtgeael (Ghent University, Ghent, Belgium)
P. Nagabhushan (University of Mysore, Mysore, India)
Bernd Neumann (Universität Hamburg, Hamburg, Germany)
Van de Weghe Nico (Ghent University, Ghent, Belgium)
Mark Nixon (University of Southampton, Southampton, UK)
Edgard Nyssen (Vrije Universiteit Brussel, Brussels, Belgium)
Daniel Ochoa (Escuela Superior Politécnica del Litoral, Guayaquil, Ecuador)
Matthias Odisio (University of Illinois at Urbana-Champaign, Urbana, USA)
Stanley Osher (UCLA, Los Angeles, USA)
Lucas Paletta (Institut für Digitale Bildverarbeitung, Graz, Austria)
Marcin Paprzycki (SWPS, Warsaw, Poland)
Jussi Parkkinen (University of Joensuu, Joensuu, Finland)
Stéphane Pateux (France Télécom, France)
Elzbieta Pekalska (Delft University, Delft, Holland)
Shmuel Peleg (The Hebrew University of Jerusalem, Jerusalem, Israel)
Rudi Penne (Karel de Grote-Hogeschool, Hoboken, Belgium)
Fernando Pereira (Instituto Superior Técnico, Lisboa, Portugal)
Herbert Peremans (University of Antwerp, Antwerp, Belgium)
Patrick Perez (IRISA / INRIA Rennes, Rennes, France)
Jean-Christophe Pesquet (Univ. Marne la Vallée, Champs sur Marne, France)
Béatrice Pesquet-Popescu (ENST, Paris, France)
Yvan Petillot (Heriot-Watt University, Edinburgh, Scotland)
Maria Petrou (Imperial College, London, UK)
Norbert Pfeifer (Institut für Geographie, Innsbruck, Austria)
Sylvie Philipp-Foliguet (ETIS, Cergy, France)
Wilfried Phillips (Ghent University, Ghent, Belgium)
Massimo Piccardi (University of Technology, Sydney, Broadway NSW, Australia)
Wojciech Pieczynski (Institut National des Télécommunications, Evry, France)
Gemma Piella (UPF, Barcelona, Spain)
Marc Pierrot-Deseilligny (Institut Géographique National, St Mandé, France)
Matti Pietikäinen (University of Oulu, Oulu, Finland)
Oliver Pietquin (Supelec, Metz, France)
Aleksandra Pizurica (Ghent University, Ghent, Belgium)
Dan Popescu (CSIRO, Sydney, Australia)
Javier Portilla (Instituto de Optica, CSIC, Madrid, Spain)
Jack-Gérard Postaire (University of Science and Technology of Lille (USTL),
Villeneuve d'Ascq Cedex, France)
Josep Prades-Nebot (Universidad Politécnica de Valencia, Valencia, Spain)
Gianni Ramponi (Trieste University, Trieste, Italy)
Thierry Ranchin (Ecole des Mines de Paris, Sophia Antipolis, France)

Alpesh Kumar Ranchordas (FMx Ltd, Crawley, UK)
Patrick Reignier (INRIA Rhône-Alpes, Saint Ismier, France)
Paolo Remagnino (Faculty of Technology, Kingston University, Surrey, UK)
Volker Rodehorst (Berlin University of Technology, Berlin, Germany)
Filip Rooms (IncGEO, Hasselt, Belgium)
Luis Salgado Álvarez de Sotomayor (Universidad Politécnica de Madrid, Madrid, Spain)
Antonio Sanz Montemayor (Universidad Rey Juan Carlos, Móstoles, Spain)
Gerald Schaefer (Nottingham Trent University, Nottingham, UK)
Peter Schelkens (Vrije Universiteit Brussel - IBBT - IMEC, Brussel, Belgium)
Paul Scheunders (University of Antwerp, Wilrijk, Belgium)
Mubarak Shah (University of Central Florida, USA)
Sheng-Wen Shih (National Chi Nan University, Puli, Taiwan)
Jan Sijbers (University of Antwerp, Wilrijk (Antwerpen), Belgium)
Athanasios Skodras (Hellenic Open University, Patras, Greece)
Olivier Stasse (CNRS/AIST, Tsukuba, Japan)
Gjenna Stippel (University of Maryland School of Medicine, Baltimore, USA)
Guan-Ming Su (University of Maryland College Park, College Park, USA)
Tomas Suk (Institute of Information Theory and Automation, Prague 8, Czech Republic)
Ming-Ting Sun (University of Washington, Seattle, USA)
Hugues Talbot (ESIEE, Noisy-le-Grand, France)
Murat Tekalp (University of Rochester, Rochester, USA)
Bruno Tellez (Université Claude Bernard Lyon, Villeurbanne, France)
Linda Tessens (Ghent University, Ghent, Belgium)
Frederic Truchetet (Université de Bourgogne, Le Creusot, France)
Filareti Tsalakanidou (Aristotle University of Thessaloniki, Thessaloniki, Greece)
Tinne Tuytelaars (KULeuven, Leuven, Belgium)
Dimitri Van De Ville (EPFL, Lausanne, Switzerland)
Gert Van de Wouwer (University of Antwerp, Wilrijk, Belgium)
Marc Van Droogenbroeck (Université de Liège, Liège, Belgium)
Ewout Vansteenkiste (Ghent University, Ghent, Belgium)
Peter Veelaert (University College Ghent, Ghent, Belgium)
Sergio Velastin (Kingston University London, UK)
Venkatraman Vidhyashankar (Cornell University, Ithaca, USA)
Jue Wang (University of Washington, USA)
Jing-Hao Xue (University of Glasgow, Glasgow, UK)
Yueming Zhu (CNRS, Villeurbanne, France)
Vladimir Zlokolica (Ghent University, Ghent, Belgium)

Table of Contents

Noise Reduction and Restoration

Directional Filtering for Upsampling According to Direction Information of the Base Layer in the JVT/SVC Codec	1
<i>Chul Keun Kim, Doug Young Suh, Gwang Hoon Park</i>	
A New Fuzzy-Based Wavelet Shrinkage Image Denoising Technique	12
<i>Stefan Schulte, Bruno Huysmans, Aleksandra Pižurica, Etienne E. Kerre, Wilfried Philips</i>	
Mathematical Models for Restoration of Baroque Paintings	24
<i>Pantaleón D. Romero, Vicente F. Candela</i>	
Motion Blur Concealment of Digital Video Using Invariant Features	35
<i>Ville Ojansivu, Janne Heikkilä</i>	
Hybrid Sigma Filter for Processing Images Corrupted by Multiplicative Noise	46
<i>Nikolay Ponomarenko, Vladimir Lukin, Karen Egiazarian, Jaakko Astola, Benoit Vozel, Kacem Chehdi</i>	
Automatic Restoration of Old Motion Picture Films Using Spatiotemporal Exemplar-Based inpainting	55
<i>Ali Gangal, Bekir Dizdaroglu</i>	
Dedicated Hardware for Real-Time Computation of Second-Order Statistical Features for High Resolution Images	67
<i>Dimitris Bariamis, Dimitris K. Iakovidis, Dimitris Maroulis</i>	
Greyscale Image Interpolation Using Mathematical Morphology	78
<i>Alessandro Ledda, Hiệp Q. Luong, Wilfried Philips, Valérie De Witte, Etienne E. Kerre</i>	
Dilation Matrices for Nonseparable Bidimensional Wavelets	91
<i>Ana Ruedin</i>	
Evolutionary Tree-Structured Filter for Impulse Noise Removal	103
<i>Nemanja I. Petrović, Vladimir S. Crnojević</i>	

Perceived Image Quality Measurement of State-of-the-Art Noise Reduction Schemes 114
Ewout Vansteenkiste, Dietrich Van der Weken, Wilfried Philips, Etienne Kerre

Multiway Filtering Applied on Hyperspectral Images 127
Nadine Renard, Salah Bourennane, Jacques Blanc-Talon

Segmentation

A Linear-Time Approach for Image Segmentation Using Graph-Cut Measures 138
Alexandre X. Falcão, Paulo A.V. Miranda, Anderson Rocha

The RIM Framework for Image Processing 150
Øyvind Ryan

A Proposal Method for Corner Detection with an Orthogonal Three-Direction Chain Code 161
Hermilo Sánchez-Cruz

A Charged Active Contour Based on Electrostatics 173
Ronghua Yang, Majid Mirmehdi, Xianguhua Xie

Comparison of Statistical and Shape-Based Approaches for Non-rigid Motion Tracking with Missing Data Using a Particle Filter 185
Abir El Abed, Séverine Dubuisson, Dominique Béréziat

An Active Contour Model Guided by LBP Distributions 197
Michalis A. Savelonas, Dimitris K. Iakovidis, Dimitris E. Maroulis, Stavros A. Karkanis

Characterizing the Lacunarity of Objects and Image Sets and Its Use as a Technique for the Analysis of Textural Patterns 208
Rafael H.C. de Melo, Evelyn de A. Vieira, Aura Conci

Adaptive Vision Leveraging Digital Retinas: Extracting Meaningful Segments 220
Nicolas Burrus, Thierry M. Bernard

A Fast Dynamic Border Linking Algorithm for Region Merging 232
Johan De Bock, Rui Pires, Patrick De Smet, Wilfried Philips

Motion Estimation and Tracking

Fast Sub-pixel Motion Estimation for H.264	242
<i>Hong Yin Lim, Ashraf A. Kassim</i>	
Temporal Error Concealment Based on Optical Flow in the H.264/AVC Standard	253
<i>Donghyung Kim, Jongho Kim, Jechang Jeong</i>	
Foreground-to-Ghost Discrimination in Single-Difference Pre-processing	263
<i>Francesco Archetti, Cristina E. Manfredotti, Vincenzina Messina, Domenico Giorgio Sorrenti</i>	
Moving Object Removal Based on Global Feature Registration.	275
<i>Soon-Yong Park, Jaekyoung Moon, Chang-Joon Park, Inho Lee</i>	
Object Tracking Using Discriminative Feature Selection.	287
<i>Bogdan Kwolek</i>	
Color-Based Multiple Agent Tracking for Wireless Image Sensor Networks	299
<i>Emre Oto, Frances Lau, Hamid Aghajan</i>	
A Fast Motion Vector Search Algorithm for Variable Blocks	311
<i>Yung-Lyul Lee, Yung-Ki Lee, HyunWook Park</i>	

Video Processing and Coding

Constrained Region-Growing and Edge Enhancement Towards Automated Semantic Video Object Segmentation	323
<i>Li Gao, Jianmin Jiang, Shuyuan Y. Yang</i>	
Spatio-temporal Composite-Features for Motion Analysis and Segmentation	332
<i>Raquel Dosil, Xosé Manuel Pardo, Xosé Ramón Fdez-Vidal, Antón García</i>	
New Intra Luma Prediction Mode in H.264/AVC Using Collocated Weighted Chroma Pixel Value	344
<i>Ik-Hwan Cho, Jung-Ho Lee, Woong-Ho Lee, Dong-Seok Jeong</i>	
Fast Mode Decision for H.264/AVC Using Mode Prediction	354
<i>Song-Hak Ri, Joern Ostermann</i>	

Performing Deblocking in Video Coding Based on Spatial-Domain Motion-Compensated Temporal Filtering.....	364
<i>Adrian Munteanu, Joeri Barbarien, Jan Cornelis, Peter Schelkens</i>	
Improving DCT-Based Coders Through Block Oriented Transforms	375
<i>Antoine Robert, Isabelle Amonou, Béatrice Pesquet-Popescu</i>	
Improvement of Conventional Deinterlacing Methods with Extrema Detection and Interpolation	384
<i>Jérôme Roussel, Pascal Bertolino, Marina Nicolas</i>	
Adaptive Macroblock Mode Selection for Reducing the Encoder Complexity in H.264	396
<i>Donghyung Kim, Jongho Kim, Jechang Jeong</i>	
Dynamic Light Field Compression Using Shared Fields and Region Blocks for Streaming Service	406
<i>Yebin Liu, Qionghai Dai, Wenli Xu, Zhihong Liao</i>	
Complexity Scalability in Motion-Compensated Wavelet-Based Video Coding	418
<i>Tom Clerckx, Adrian Munteanu, Jan Cornelis, Peter Schelkens</i>	
Spatial Error Concealment with Low Complexity in the H.264 Standard	431
<i>Donghyung Kim, Seungjong Kim, Jechang Jeong</i>	
A Real-Time Content Adaptation Framework for Exploiting ROI Scalability in H.264/AVC	442
<i>Peter Lambert, Davy De Schrijver, Davy Van Deursen, Wesley De Neve, Yves Dhondt, Rik Van de Walle</i>	
Complexity Reduction Algorithm for Intra Mode Selection in H.264/AVC Video Coding.....	454
<i>Jongho Kim, Donghyung Kim, Jechang Jeong</i>	
Simple and Effective Filter to Remove Corner Outlier Artifacts in Highly Compressed Video	466
<i>Jongho Kim, Donghyung Kim, Jechang Jeong</i>	
Content-Based Model Template Adaptation and Real-Time System for Behavior Interpretation in Sports Video	474
<i>Jungong Han, Peter H.N. de With</i>	
New Approach to Wireless Video Compression with Low Complexity	485
<i>Gangyi Jiang, Zhipeng Jin, Mei Yu, Tae-Young Choi</i>	

Fast Multi-view Disparity Estimation for Multi-view Video Systems	493
<i>Gangyi Jiang, Mei Yu, Feng Shao, You Yang, Haitao Dong</i>	
AddCanny: Edge Detector for Video Processing	501
<i>Luis Antón-Canalís, Mario Hernández-Tejera, Elena Sánchez-Nielsen</i>	
Video-Based Facial Expression Hallucination: A Two-Level Hierarchical Fusion Approach	513
<i>Jian Zhang, Yueting Zhuang, Fei Wu</i>	
Blue Sky Detection for Picture Quality Enhancement	522
<i>Bahman Zafarifar, Peter H.N. de With</i>	
Requantization Transcoding in Pixel and Frequency Domain for Intra 16x16 in H.264/AVC	533
<i>Jan De Cock, Stijn Notebaert, Peter Lambert, Davy De Schrijver, Rik Van de Walle</i>	
Motion-Compensated Deinterlacing Using Edge Information	545
<i>Taeuk Jeong, Chulhee Lee</i>	
Video Enhancement for Underwater Exploration Using Forward Looking Sonar	554
<i>Kio Kim, Nicola Neretti, Nathan Intrator</i>	
Camera Calibration, Image Registration and Stereo Matching	
Optimal Parameters Selection for Non-parametric Image Registration Methods	564
<i>Jorge Larrey-Ruiz, Juan Morales-Sánchez</i>	
Camera Calibration from a Single Frame of Planar Pattern	576
<i>Jianhua Wang, Fanhuai Shi, Jing Zhang, Yuncai Liu</i>	
Stereo Matching Using Scanline Disparity Discontinuity Optimization	588
<i>Ho Yub Jung, Kyoung Mu Lee, Sang Uk Lee</i>	
A New Stereo Matching Model Using Visibility Constraint Based on Disparity Consistency	598
<i>Ju Yong Chang, Kyoung Mu Lee, Sang Uk Lee</i>	

Refine Stereo Correspondence Using Bayesian Network and Dynamic Programming on a Color Based Minimal Span Tree 610
Naveed Iqbal Rao, Huijun Di, GuangYou Xu

Estimation of Rotation Parameters from Blurred Image 620
Qian Li, Shi-gang Wang

Hierarchical Stereo Matching: From Foreground to Background 632
Kai Zhang, Yuzhou Wang, Guoping Wang

Biometrics and Security

Gabor Feature Based Face Recognition Using Supervised Locality Preserving Projection 644
Zhonglong Zheng, Jianmin Zhao, Jie Yang

Alternative Fuzzy Clustering Algorithms with L1-Norm and Covariance Matrix 654
Miin-Shen Yang, Wen-Liang Hung, Tsiung-Iou Chung

A Statistical Approach for Ownership Identification of Digital Images 666
Ching-Sheng Hsu, Shu-Fen Tu, Young-Chang Hou

Rigid and Non-rigid Face Motion Tracking by Aligning Texture Maps and Stereo-Based 3D Models 675
Fadi Dornaika, Angel D. Sappa

Curve Mapping Based Illumination Adjustment for Face Detection 687
Xiaoyue Jiang, Tuo Zhao, Rongchun Zhao

Common Image Method(Null Space + 2DPCAs) for Face Recognition 699
Hae Jong Seo, Young Kyung Park, Joong Kyu Kim

Discrete Choice Models for Static Facial Expression Recognition 710
Gianluca Antonini, Matteo Sorci, Michel Bierlaire, Jean-Philippe Thiran

Scalable and Channel-Adaptive Unequal Error Protection of Images with LDPC Codes 722
Adrian Munteanu, Maryse R. Stoufs, Jan Cornelis, Peter Schelkens

Robust Analysis of Silhouettes by Morphological Size Distributions	734
<i>Olivier Barnich, Sébastien Jodogne, Marc Van Droogenbroeck</i>	
Enhanced Watermarking Scheme Based on Texture Analysis	746
<i>Ivan O. Lopes, Celia A.Z. Barcelos, Marcos A. Batista, Anselmo M. Silva</i>	
A Robust Watermarking Algorithm Using Attack Pattern Analysis	757
<i>Dong Eun Lee, Taekyung Kim, Seongwon Lee, Joonki Paik</i>	
Probability Approximation Using Best-Tree Distribution for Skin Detection	767
<i>Sanaa El Fkihi, Mohamed Daoudi, Driss Aboutajdine</i>	
Fusion Method of Fingerprint Quality Evaluation: From the Local Gabor Feature to the Global Spatial-Frequency Structures	776
<i>Decong Yu, Lihong Ma, Hanqing Lu, Zhiqing Chen</i>	
3D Face Recognition Based on Non-iterative Registration and Single B-Spline Patch Modelling Techniques	786
<i>Yi Song, Li Bai</i>	
Automatic Denoising of 2D Color Face Images Using Recursive PCA Reconstruction	799
<i>Hyun Park, Young Shik Moon</i>	
Facial Analysis and Synthesis Scheme	810
<i>Ilse Ravyse, Hichem Sahli</i>	
Medical Imaging	
Detection of Pathological Cells in Phase Contrast Cytological Images	821
<i>Marcin Smereka, Grzegorz Glab</i>	
Water Flow Based Complex Feature Extraction	833
<i>Xin U Liu, Mark S Nixon</i>	
Seeded Region Merging Based on Gradient Vector Flow for Image Segmentation	846
<i>Yuan He, Yupin Luo, Dongcheng Hu</i>	
System for Reading Braille Embossed on Beverage Can Lids for Authentication	855
<i>Trine Kirkhus, Jens T Thielemann, Britta Fismen, Henrik Schumann-Olsen, Ronald Sivertsen, Mats Carlin</i>	

Leukocyte Segmentation in Blood Smear Images Using Region-Based Active Contours 867
Seongeun Eom, Seungjun Kim, Vladimir Shin, Byungha Ahn

Multiresolution Lossy-to-Lossless Coding of MRI Objects 877
Habibollah Danyali, Alfred Mertins

A Novel Fuzzy Segmentation Approach for Brain MRI 887
Gang Yu, Changguo Wang, Hongmei Zhang, Yuxiang Yang, Zhengzhong Bian

Extrema Temporal Chaining: A New Method for Computing the 2D-Displacement Field of the Heart from Tagged MRI 897
Jean-Pascal Jacob, Corinne Vachier, Jean-Michel Morel, Jean-Luc Daire, Jean-Noel Hyacinthe, Jean-Paul Vallée

Data Fusion and Fuzzy Spatial Relationships for Locating Deep Brain Stimulation Targets in Magnetic Resonance Images 909
Alice Villéger, Lemlih Ouchchane, Jean-Jacques Lemaire, Jean-Yves Boire

Robust Tracking of Migrating Cells Using Four-Color Level Set Segmentation 920
Sumit K. Nath, Filiz Bunyak, Kannappan Palaniappan

Image Retrieval and Image Understanding

Robust Visual Identifier for Cropped Natural Photos 933
Ik-Hwan Cho, A-Young Cho, Hae-Kwang Kim, Weon-Geun Oh, Dong-Seok Jeong

Affine Epipolar Direction from Two Views of a Planar Contour 944
Maria Alberich-Carramiñana, Guillem Alenyà, Juan Andrade-Cetto, Elisa Martínez, Carme Torras

Toward Visually Inferring the Underlying Causal Mechanism in a Traffic-Light-Controlled Crossroads 956
Joaquín Salas, Sandra Canchola, Pedro Martínez, Hugo Jiménez, Reynaldo C. Pless

Computer Vision Based Travel Aid for the Blind Crossing Roads 966
Tadayoshi Shioyama

A Novel Stochastic Attributed Relational Graph Matching Based on Relation Vector Space Analysis	978
<i>Bo Gun Park, Kyoung Mu Lee, Sang Uk Lee</i>	
A New Similarity Measure for Random Signatures: Perceptually Modified Hausdorff Distance	990
<i>Bo Gun Park, Kyoung Mu Lee, Sang Uk Lee</i>	
Tracking of Linear Appearance Models Using Second Order Minimization	1002
<i>Jose Gonzalez-Mora, Nicolas Guil, Emilio L. Zapata</i>	
Visibility of Point Clouds and Mapping of Unknown Environments	1014
<i>Yanina Landa, Richard Tsai, Li-Tien Cheng</i>	
Adjustment for Discrepancies Between ALS Data Strips Using a Contour Tree Algorithm	1026
<i>Dongyeob Han, Jaebin Lee, Yongil Kim, Kiyun Yu</i>	
Visual Bootstrapping for Unsupervised Symbol Grounding	1037
<i>Josef Kittler, Mikhail Shevchenko, David Windridge</i>	
A 3D Model Acquisition System Based on a Sequence of Projected Level Curves	1047
<i>Huei-Yung Lin, Ming-Liang Wang, Ping-Hsiu Yu</i>	
Scale Invariant Robust Registration of 3D-Point Data and a Triangle Mesh by Global Optimization	1059
<i>Onay Urfalioğlu, Patrick A. Mikulastik, Ivo Stegmann</i>	
Fast Hough Transform Based on 3D Image Space Division	1071
<i>Witold Zorski</i>	
Context-Based Scene Recognition Using Bayesian Networks with Scale-Invariant Feature Transform.....	1080
<i>Seung-Bin Im, Sung-Bae Cho</i>	
A Portable and Low-Cost E-Learning Video Capture System	1088
<i>Richard Yi Da Xu</i>	
On Building Omnidirectional Image Signatures Using Haar Invariant Features: Application to the Localization of Robots	1099
<i>Cyril Charron, Ouidad Labbani-Igbida, El Mustapha Mouaddib</i>	
Accurate 3D Structure Measurements from Two Uncalibrated Views	1111
<i>Benjamin Albouy, Emilie Koenig, Sylvie Treuillet, Yves Lucas</i>	

A Fast Offline Building Recognition Application on a Mobile Telephone 1122
Nikolaj J.C. Groeneweg, Bastiaan de Groot, Arvid H.R. Halma, Bernardo R. Quiroga, Maarten Tromp, Frans C.A. Groen

Adaptive Learning Procedure for a Network of Spiking Neurons and Visual Pattern Recognition 1133
Simej Gomes Wysoski, Lubica Benuskova, Nikola Kasabov

Interactive Learning of Scene Context Extractor Using Combination of Bayesian Network and Logic Network 1143
Keum-Sung Hwang, Sung-Bae Cho

Classification and Recognition

Adaptative Road Lanes Detection and Classification 1151
Juan M. Collado, Cristina Hilario, Arturo de la Escalera, Jose M. Armingol

An Approach to the Recognition of Informational Traffic Signs Based on 2-D Homography and SVMs 1163
Amelio Vázquez-Reina, Roberto Javier López-Sastre, Philip Siegmann, Sergio Lafuente-Arroyo, Hilario Gómez-Moreno

On Using a Dissimilarity Representation Method to Solve the Small Sample Size Problem for Face Recognition 1174
Sang-Woon Kim

A Comparison of Nearest Neighbor Search Algorithms for Generic Object Recognition 1186
Ferid Bajramovic, Frank Mattern, Nicholas Butko, Joachim Denzler

Non Orthogonal Component Analysis: Application to Anomaly Detection 1198
Jean-Michel Gaucel, Mireille Guillaume, Salah Bourennane

A Rough Set Approach to Video Genre Classification 1210
Wengang Cheng, Chang'an Liu, Xingbo Wang

Author Index 1221

Directional Filtering for Upsampling According to Direction Information of the Base Layer in the JVT/SVC Codec

Chul Keun Kim, Doug Young Suh, and Gwang Hoon Park

Media Laboratory, Kyunghee Univ. Seochunri, Giheungeuop, Younginsi, Kyungido,
Korea

{chulkeun, suh, ghpark}@khu.ac.kr
<http://medialab.khu.ac.kr>

Abstract. When the reconstructed image of the base layer is up-sampled for decoding the enhancement layer of spatial scalability, direction information derived during decoding the spatially lower layer is used. That is direction information used for Intra prediction which is used for up-sampling again. In most cases, it shows 0.1-0.5dB quality improvement in images up-sampled by using directional filtering compared to those up-sampled conventionally. The same interpolation algorithm should be used in both the encoder and decoder.

1 Introduction

In a SVC, up-sampling of video sequences is an important process for spatial scalability. In order to re-use the lower layer information, the base layer images should be up-sampled to the size of the higher layer images. The rate-distortion performance of spatial scalable video coding directly depends on the up-sampling methods used in the encoder and decoder. Currently for inter blocks and intra blocks, 2-tap filter and 6-tap filter are used, respectively. These filters achieve up-sampling by interpolating pixels horizontally and vertically. By using this method, horizontal and vertical high frequency is well preserved. However, this method is not efficient because characteristics of blocks are different. Intra prediction in H.264 uses the minimum SAE (Sum of Absolute Error) method out of the 9 prediction modes. Through prediction we can see that this mode shows direction characteristics of blocks. This paper proposes to use directional filtering according to direction information used Intra prediction in the baser layer. It includes directional filtering method for any direction out of 8 directions of intra prediction. Section 2 describes Characteristics of up-sampling in JVT (Joint Video Team) / SVC (Scalable Video Coding).[6] Section 3 describes why directional up-sampling is needed. Section 4 proposes directional up-sampling method. It is followed by Section 5 which describes experiments for performance analysis of directional filter. Section 6 concludes this paper.

2 Characteristics of Up-Sampling Filter in JSVM

To generate a multi-resolution input for scalable video coding, the down-sampling of (1) is used in [6]. For the reuse of the lower layer at the higher layer, up-sampling filter of (2) is used in [6].

$$H_{down} = \frac{[2, 0, -4, -3, 5, 19, 26, 19, 5, -3, -4, 0, 2]}{64} \tag{1}$$

$$H_{up} = \frac{[1, -5, 20, 20, -5, 1]}{32} \tag{2}$$

Fig.1 shows characteristics of the 6-tap filter compared to the ideal LPF. Area A shows loss of energy while area B shows imperfect cut-off. These two areas result in distortion in up-sampling.

Fig.2.(a) shows the 2D response of an ideal LPF. It can be seen that the high frequency is removed and the low frequency is left. Fig.2.(b) and fig.2.(c) show the frequency response of a 6-tap filter, where due to area A and B, high frequency is left

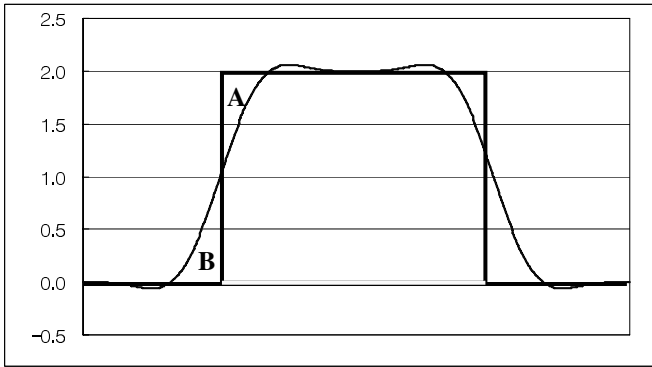


Fig. 1. Characteristics of a 6-tap filter and a ideal LPF

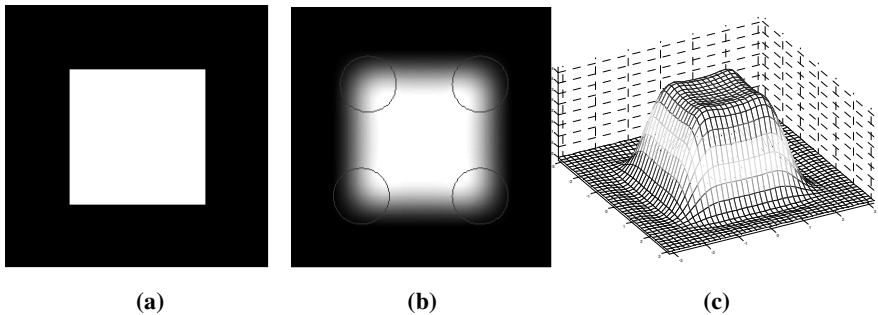


Fig. 2. (a) Frequency response of a ideal LPF, (b) and (c) frequency response of a 6-tap filter

and low frequency is removed, distortion occurs. It can be seen the most distortion occurs where area A and B is applied horizontal and vertically. We propose the use of directional filter to reduce the distortion .

3 Necessity of Directional Upsampling

Fig.3. (a)~(c) shows vertical(90°), horizontal(0°), diagonal(45°) features and 2D frequency response. Fig.3.(a)~(c) show concentrated distribution horizontally, vertically and diagonally respectively.

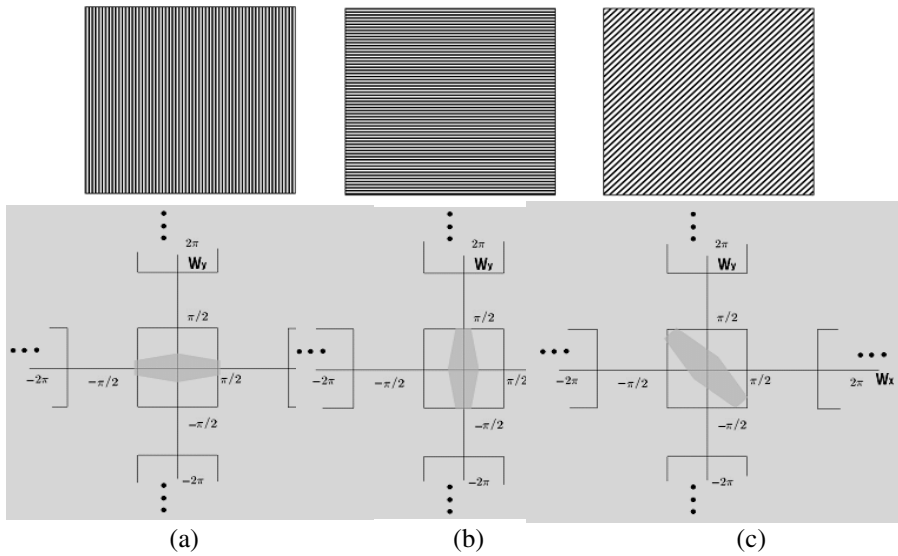


Fig. 3. (a) vertical(90°) features and 2D frequency response, (b) horizontal(0°) features and 2D frequency response, (c) diagonal(45°) features and 2D frequency response

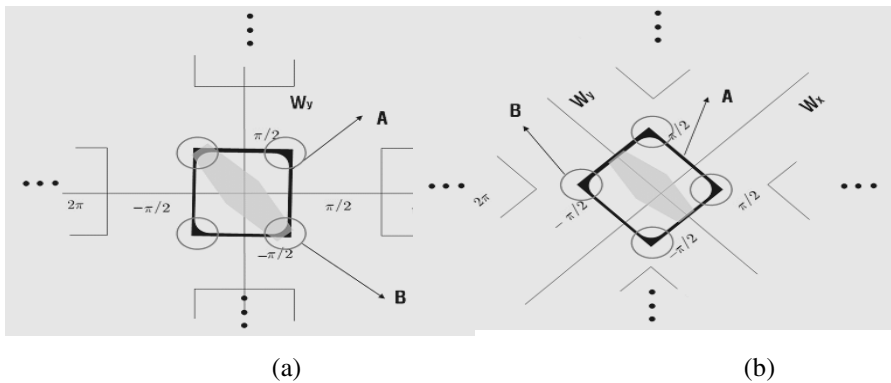


Fig. 4. (a) Conventional filter (b) Proposed directional filter

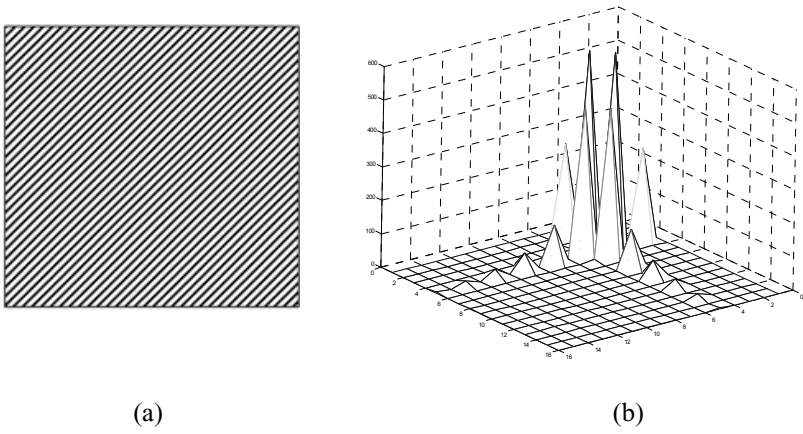


Fig. 5. (a) Example of diagonal image (b) DCT representation of diagonal image

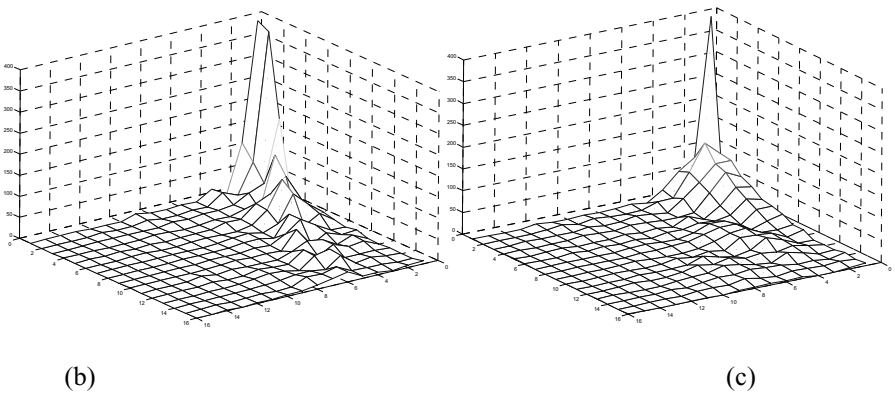
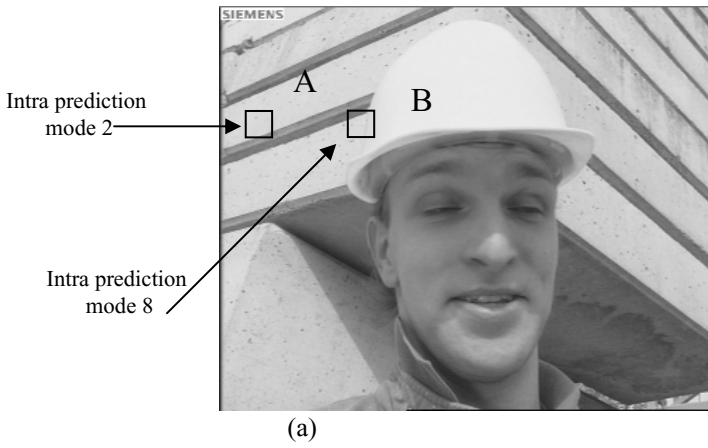


Fig. 6. DCT representation of blocks with directional information

Fig.4.(a) shows the example when the features of a image are like fig.3.(c) and a conventional up-sampled filter is used. The black area in fig.4 represents the distorted area when a 6-tap filter is used and A and B show the extreme areas. In the conventional up-sampling filter, much low frequency is distorted in area. This case the directional filter is used to up-sample and we can avoid area A and B .fig.4.(b) shows that this has less distortion.

Fig.5.(a) show a -45° feature image and fig.7.(b) is its DCT representation. Normally, the distribution is near the DC component, but high AC components can be seen

Fig.6 shows in the DCT domain of which contains blocks with directional information and the other without it. Fig.6.(a) show block A in intra prediction mode 2 which shows DC characteristics and block B is intra prediction mode 8 with 22.5° characteristics. Fig.6 (b) represents the DCT coefficient distribution of block B and Fig.6.(c) shows DCT coefficient distribution of block B. The DCT coefficients of block A show concentration near the DC while Block B shows it is distributed towards it directional characteristics at AC.

4 Method of Directional Upsampling

Currently, the low pass filter shown in fig.7.(a) and fig.7.(b) is used for vertical and horizontal 1:2 interpolation of the whole picture. We propose a diagonal 1:2

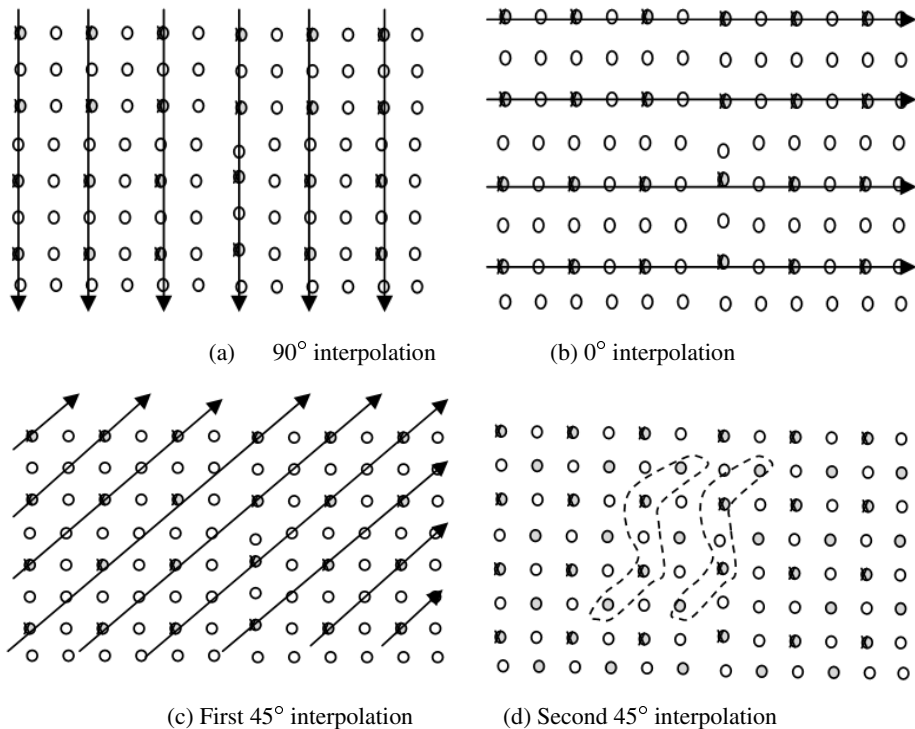
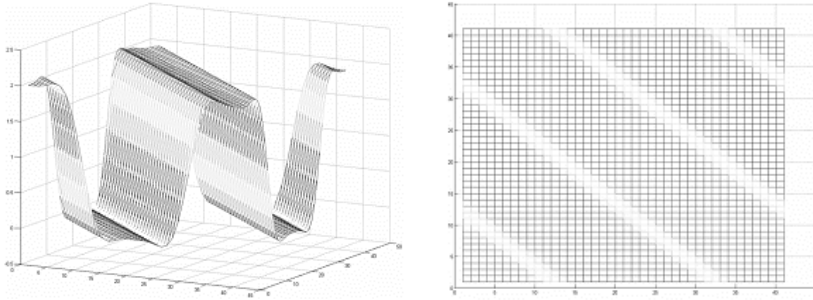
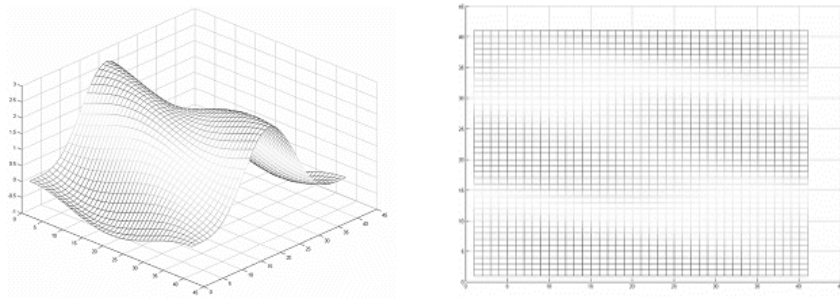


Fig. 7. Directional Interpolation (gray: non-zero sample, white : zero-pad)



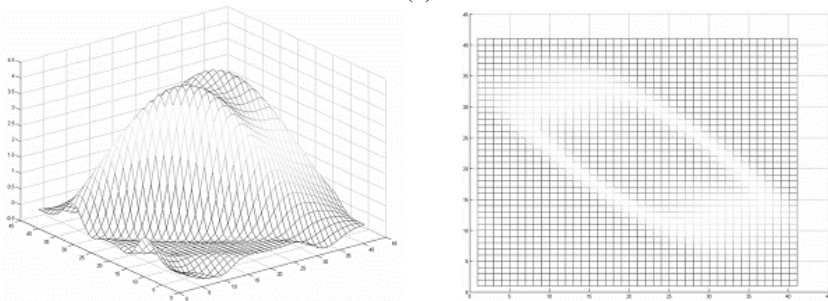
$$H_{d1}(\omega_x, \omega_y) = (32 + 40\cos(\omega_x + \omega_y) - 10\cos(3\omega_x + 3\omega_y) + 2\cos(5\omega_x + 5\omega_y)) / 32$$

(a)



$$H_{d2}(\omega_x, \omega_y) = (30 + 40\cos_y - 10\cos(\omega_x + 2\omega_y)) / 30$$

(b)



$$H_d(\omega_x, \omega_y) = H_{d1}(\omega_x, \omega_y)H_{d2}(\omega_x, \omega_y)$$

(c)

Fig. 8. (a) and (b) each represent the fourier transform and frequency response of the first and second 45° interpolation. Fig.9.(c) is the sequential application of these filters represented in fourier transform and frequency response of the directional up-sampling. It can be seen in fig.9.(c) that this filter has a -45° characteristic and can be applied to intra prediction mode 3.

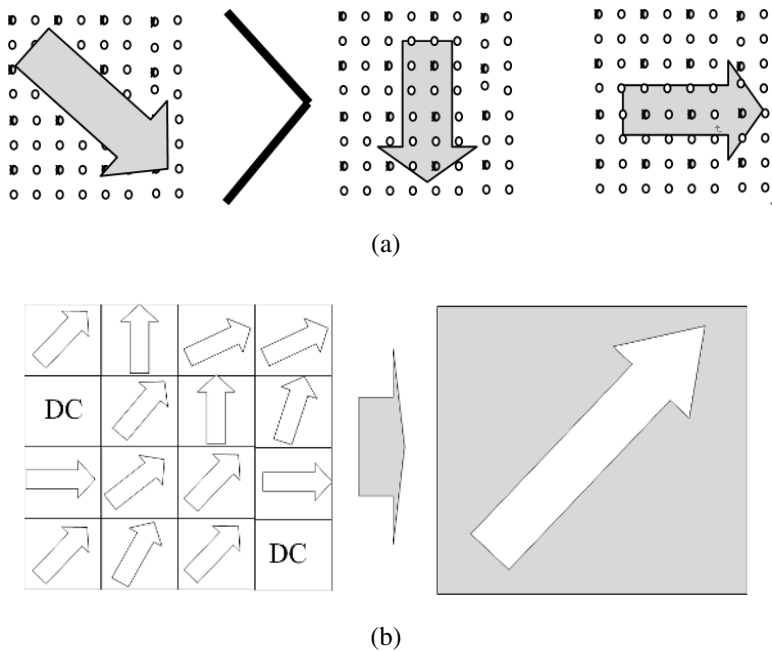


Fig. 9. Our proposed method to reduce bad effect of unsuitable directional filtering is used only when a diagonal characteristic is larger than the other directional characteristics

Table 1. Combination of directional interpolation for 8 direction modes

Modes	0, 1	3	4	5	6	7	8
Angles	90° , 0°	45°	-45°	-67.5°	-22.5°	67.5°	22.5°
1 st interpolation	0°	45°	-45°	-45°	-45°	45°	45°
2 nd interpolation	90°	45°	-45°	90°	0°	90°	0°

interpolation by using the low pass filter as shown in fig.7.(c) and fig.7.(d). Before applying the directional filter, the use must be determined by the features of the block and the intra prediction mode. By combining three interpolation directions such as vertical(90°), horizontal(0°), and diagonal(45°) interpolation, we could interpolate in any direction.

For example, 45° and -45° interpolation are used for mode 3(45°) while 45° and 90° interpolation is used for mode 7(67.5°). Table.1 shows a choice of directional interpolation for 8 different directions.

Interpolation would have a bad effect by unsuitable directional filtering cause of the fact that Intra prediction mode does not show the directional characteristics of block. There are 16 4×4 blocks in a macroblock. When the directions of the blocks differ, the

bad effects increase. In order to reduce computational complexity and bad effects, a direction for each macroblock is determined from 16 blocks.

4.1 Ratio of Macroblocks with Respect to Directions

Fig.10 shows the ratio of macroblocks with respect to directions of the bitrate of an image. It can be seen that the Ratio of macroblocks with respect to direction changes with no special characteristic respecting to bitrate. In all images, ratio of macroblocks with vertical or horizontal direction is less than 50%. They are interpolated by using conventional vertical and horizontal 1:2 interpolation filtering. The other macroblocks are interpolated by using the proposed directional 1:2 interpolation filtering.

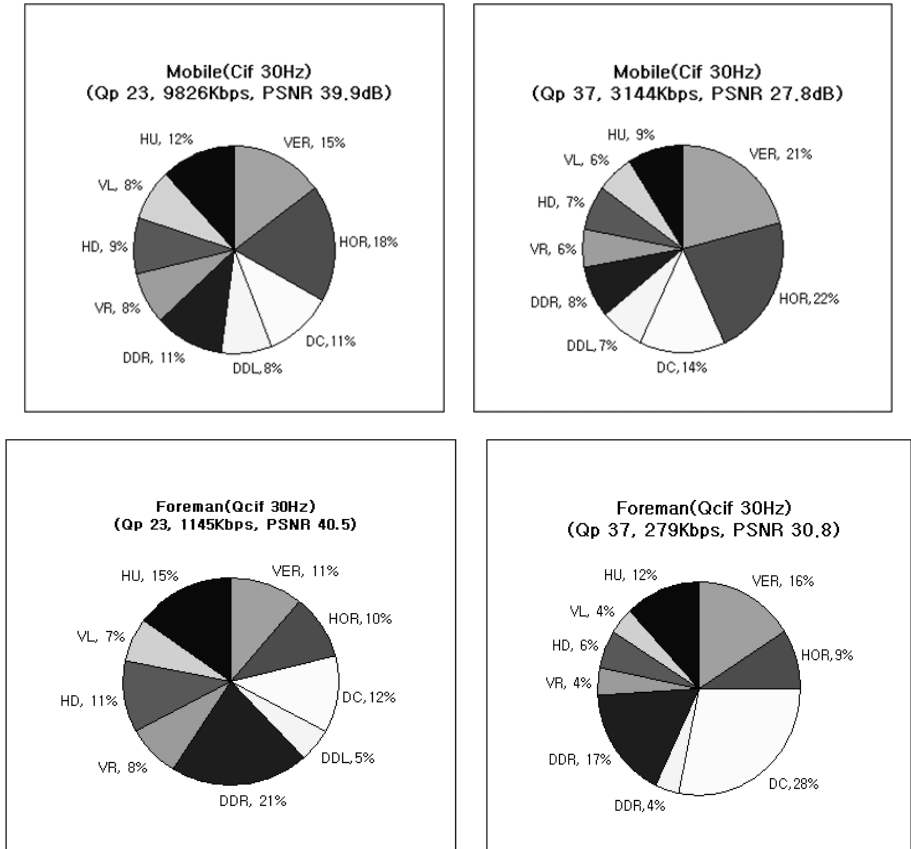


Fig. 10. Ratio of macroblocks with respect to directions

4.2 Subjective Improvement and Objective Improvement

Fig.11 shows a region where directional interpolation improves subjective quality substantially. In this case 22.5° interpolation is applied. Block B is area A upsampled with the existing filter and block C is area A upsampled with a 22.5 degree directional filter.

By comparing area D of block B and area E of block C, it can be seen that the edges of area E at a 22.5 direction is more active. The PSNR of block C is 31.7dB, which is 0.4dB higher than the 31.3dB of block B.

Fig. 12 shows the R-D curves of the proposed method and current method. The proposed interpolation has a higher PSNR at about 0.1dB~0.5dB. The higher PSNR of the decoded image, the more effectivity of proposed method. This is caused by the fact that intra prediction mode show the directionality of image signal at high quality. In the case of high bitrate encoded bitstream, the quality of decoded image is increased up to 0.5dB.

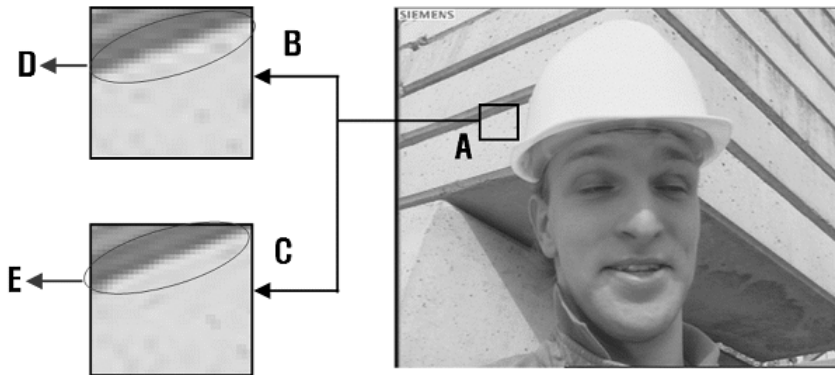
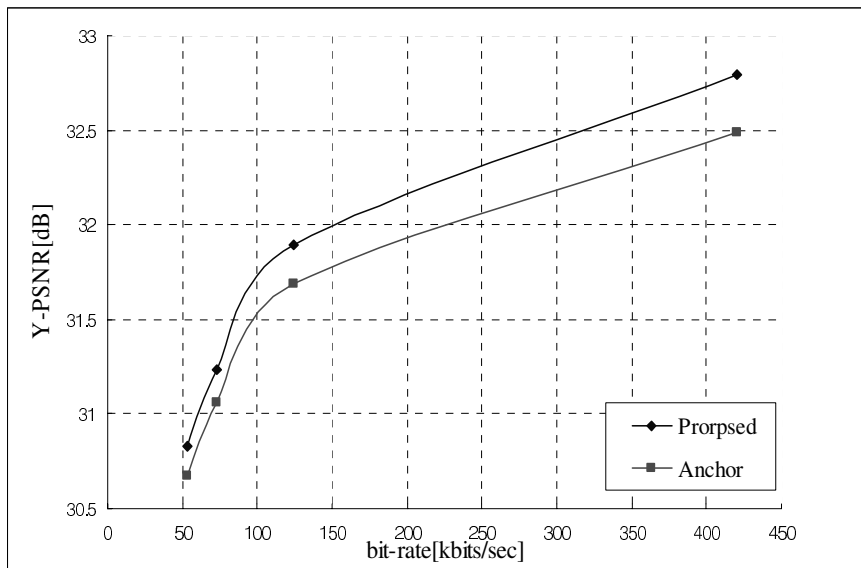
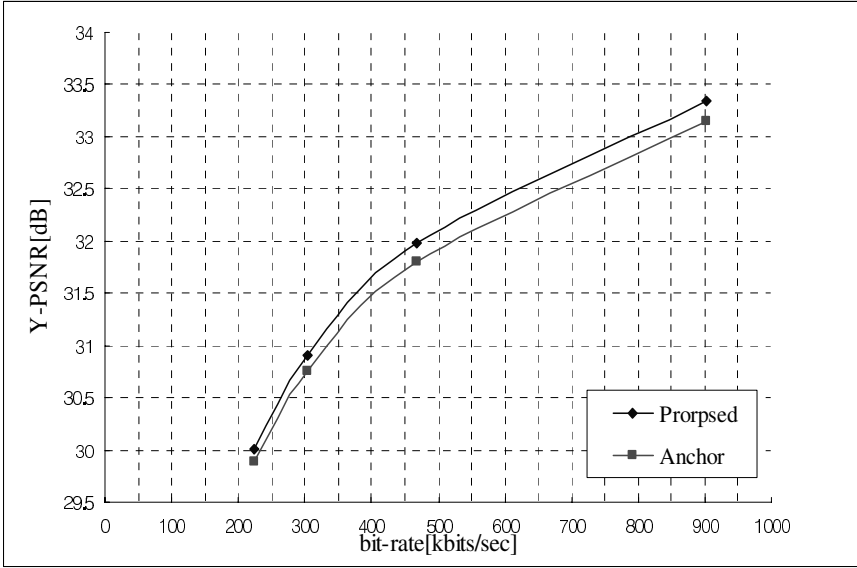


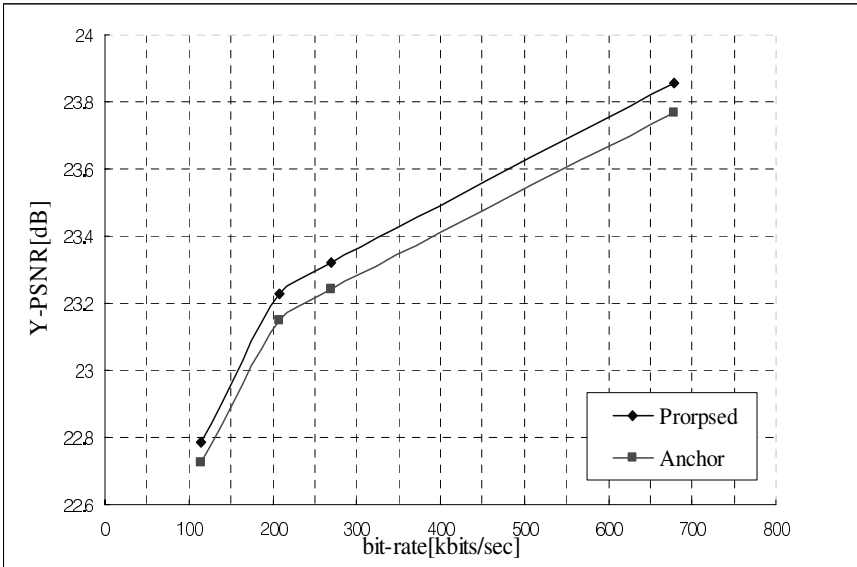
Fig. 11. D : interpolated by conventional filtering, E : interpolated by proposed directional filtering



(a)Foreman



(b)Football



(c)Mobile

Fig. 12. R-D curves of Proposed method and current method

5 Conclusions

This paper suggests using direction information of intra prediction again in up-sampling for spatial scalable coding. This paper, also, provides directional interpolation by introducing a method to apply the current low pass filter used in vertical and horizontal 1:2 interpolation to diagonal 1:2 interpolation. The same up-sampling procedure should be applied to both encoding and decoding. Experiments show that more than 50% macroblocks are so inclined that the proposed directional interpolation could be performed. No syntax change is required, but only a change in up-sampling process is required. Logical complexity is increased since interpolation is performed macroblock by macroblock while currently it is performed on the whole picture. The computation amount, however, remains almost the same because the computation amount per pixel is the same as the current. Up-sampled images by using bad effect excluded proposed directional filter are better by 0.1-0.5dB in PSNR, which results in quality improvement of decoded images.

Acknowledgements

This work is supported by Korea Science & Engineering Foundation through the National Research Laboratory(NRL) program (Grant 2005-01363).

References

1. Ilhong Shin and HyunWook Park :Down/up-sample methods for spatial scalability of SVC , JVT-O024 , 15th JVT Meeting
2. Mathias Wien : Variable Block-Size Transforms for H.264/AVC, IEEE TRANSACTIONS ON CIRCUITS AND SYSTEMS FOR VIDEO TECHNOLOGY, VOL. 13, NO. 7, JULY 2003
3. Heiko Schwarz, Tobias Hinz, Detlev Marpe, and Thomas Wiegand :Constrained Inter-Layer Prediction forSingle-Loop Decoding in Spatial Scalability, Proc. ICIP 2005, Genova, Italy, September 11-14, 2005.
4. H. Schwarz, D. Marpe, and T. Wiegand: MCTF and Scalability Extension of H.264/AVC, Proc. of PCS 2004, San Francisco, CA,USA, Dec. 2004.
5. SVM 3.0 Software, ISO/IEC JTC1/ SC29/ WG11 N6717, October 2004.
6. Joint Video Team (JVT) of ISO/IEC MPEG & ITU-T VCEG, 'Joint Scalable Video Model JSVM-2', JVT-O202, Busan, Korea, April, 2005

A New Fuzzy-Based Wavelet Shrinkage Image Denoising Technique

Stefan Schulte¹, Bruno Huysmans², Aleksandra Pižurica^{2,*},
Etienne E. Kerre¹, and Wilfried Philips²

¹ Ghent University, Department of Applied Mathematics and Computer Science,
Krijgslaan 281 (Building S9), 9000 Gent, Belgium

² Ghent University, Dept. of Telecommunications and Information Processing
(TELIN), IPI, Sint-Pietersnieuwstraat 41, 9000 Gent, Belgium

`Stefan.Schulte@UGent.be`

Abstract. This paper focuses on fuzzy image denoising techniques. In particular, we investigate the usage of fuzzy set theory in the domain of image enhancement using wavelet thresholding. We propose a simple but efficient new fuzzy wavelet shrinkage method, which can be seen as a fuzzy variant of a recently published probabilistic shrinkage method [1] for reducing adaptive Gaussian noise from digital greyscale images. Experimental results show that the proposed method can efficiently and rapidly remove additive Gaussian noise from digital greyscale images. Numerical and visual observations show that the performance of the proposed method outperforms current fuzzy non-wavelet methods and is comparable with some recent but more complex wavelets methods. We also illustrate the main differences between this version and the probabilistic version and show the main improvements in comparison to it.

1 Introduction

In general, image denoising imposes a compromise between noise reduction on the one hand and preserving significant image details on the other hand. To achieve a good performance, a noise reduction algorithm should adapt itself to the spatial context. The wavelet transform [2] significantly facilitates the construction of such spatially adaptive algorithms, due to its energy compaction property: it compresses the essential information into a few large coefficients which represent the image details along several resolution scales.

Typical wavelet based denoising methods consist of three steps: (i) compute the discrete wavelet transform (DWT) or a non-decimated wavelet transform, (ii) remove noise from the wavelet coefficients and (iii) reconstruct the enhanced image by using the inverse wavelet transformation. Due to the linearity of the wavelet transform, additive noise in the image domain remains additive in the transform domain as well. If $w_{s,d}(i, j)$ and $y_{s,d}(i, j)$ denote the noisy, respectively

* “A. Pižurica is a postdoctoral researcher of the Fund for the Scientific Research in Flanders (FWO) Belgium”.

the noise-free wavelet coefficients of scale s and orientation d then we can model the additive noise in the transform domain as:

$$w_{s,d}(i, j) = y_{s,d}(i, j) + n_{s,d}(i, j) \quad (1)$$

where $n_{s,d}(i, j)$ is the corresponding noise component. In this paper we restrict ourselves to additive Gaussian noise.

The second step in the wavelet denoising procedure usually consists of *shrinking* the wavelet coefficients: the coefficients that contain primarily noise should be reduced to negligible values, while the ones containing a significant noise free component should be reduced less. A common shrinkage approach is the application of simple *thresholding* nonlinearities to the empirical wavelet coefficients [3,4,5]: if the coefficient's magnitude is below the threshold T it is reduced to zero, otherwise it is kept or modified. Shrinkage estimators can also result from a *Bayesian approach*, in which a prior distribution of the noise-free data (e.g., Laplacian [6], generalized Gaussian [7,8,9], Gaussian Scale Mixture [10]) is integrated in the denoising scheme. The simplest Bayesian methods assume statistically independent data and rely on marginal statistics only [7,8,11,12].

However, algorithms that exploit the different kinds of dependencies between the wavelet coefficients can result in better denoising performance, compared with the ones derived using an independence assumption. The wavelet coefficients are statistically dependent mainly due to two properties of the wavelet transform of natural images: (1) large coefficients will propagate across the scales (interscale dependencies), and (2) if a coefficient is large/small, some of the neighbouring coefficients are also likely to be large/small (intrascale dependencies).

Recently, non-Gaussian bivariate distributions capturing the *interscale dependency* were proposed[13], and corresponding nonlinear shrinking functions were derived from these distributions using Bayesian estimation theory. Interscale dependencies among the wavelet coefficients are also often modelled with Hidden Markov Trees (HMT)[14,15]. Related methods [9,16,17] use Markov Random Field (MRF) models for capturing intrascale (spatial) dependencies among the wavelet coefficients. It has been proved useful to combine the first order statistical properties of the coefficient magnitudes and their evolution across scales within a joint statistical distribution model [9].

Many other techniques combine inter- and intrascale dependencies. For example, denoising methods based on Gaussian Scale Mixture models, often employ the neighbouring coefficients on the same and adjacent scales [10]. Locally adaptive window-based methods [1,18] are highly performant despite their simplicity. Local contextual HMT models have been developed, which capture both interscale and intrascale information[19,20].

If a certain wavelet coefficient and its neighbouring coefficients are small enough we know that this coefficient is noisy for almost sure and should be put equal to zero. Coefficients above a certain threshold contain the most important image structures and should not be reduced, but coefficients with values around the threshold contain both noise and signals of interest. A good threshold is generally chosen so that most coefficients below the threshold are noise

and values above the threshold are signals of interest. In such situation it can be advantageous to use fuzzy set theory as kind of soft-threshold method. Fuzzy set theory is a mathematical extension of the binary set theory.

Fuzzy set theory and fuzzy logic [21] offer us powerful tools to represent and process human knowledge represented as fuzzy if-then rules. Fuzzy image processing [22] has three main stages: (i) image fuzzification, (ii) modification of membership values and (iii) image defuzzification. The fuzzification and defuzzification steps are due to the fact that we do not yet possess fuzzy hardware. Therefore, the coding of image data (fuzzification) and decoding of the results (defuzzification) are steps that make it possible to process images with fuzzy techniques. The main power of fuzzy image processing lies in the second step (modification of membership values). After the image data is transformed from input plane to the membership plane (fuzzification), appropriate fuzzy techniques modify the membership values. This can be a fuzzy clustering, a fuzzy rule-based approach, a fuzzy integration approach, etc.

The main advantages of the new method are: (i) the complexity of the method is much lower than the probabilistic one [1] (which results in a lower execution time), (ii) we do not lose any noise reduction performance and (iii) by adding new fuzzy rules it should be easily extendable to incorporate other information as well (e.g. interscale or interband information), to further improve the noise reduction performance (future work).

The paper is structured as follows: In section 2 we discuss the proposed fuzzy shrinkage method. Experimental results are presented in section 3 and section 4 concludes the paper.

2 Fuzzy Shrinkage Method

We develop a novel fuzzy wavelet shrinkage method, which is a fuzzy-logic variant of the recent *ProbShrink* method of [1]. The method of [1] defines for each coefficient $w_{s,d}(i, j)$ two hypotheses: H_1 : signal of interest present ($|y_{s,d}(i, j)| > \sigma$) and H_0 : signal of interest absent ($|y_{s,d}(i, j)| \leq \sigma$). The method was named *ProbShrink* because it shrinks each coefficient according to probability that the coefficient presents a signal of interest given its value $w_{s,d}(i, j)$ and given a local spatial activity indicator $x_{s,d}(i, j)$ as follows: $\hat{y}_{s,d}(i, j) = P(H_1|w_{s,d}(i, j), x_{s,d}(i, j)) w_{s,d}(i, j)$. The local spatial activity indicator was defined as the average magnitude of the surrounding wavelet coefficients within a local window. In our notation, this is:

$$x_{s,d}(i, j) = \frac{\left(\sum_{k=-K}^K \sum_{l=-K}^K |w_{s,d}(i+k, j+l)| \right) - |w_{s,d}(i, j)|}{(2K+1)^2 - 1} \quad (2)$$

The method of [1] proceeds by estimating the conditional probability density functions of $w_{s,d}(i, j)$ and $x_{s,d}(i, j)$ given H_1 and given H_0 and by using the corresponding likelihood ratios: $\xi(w_{s,d}(i, j)) = p(w_{s,d}(i, j)|H_1)/p(w_{s,d}(i, j)|H_0)$

and $\eta(x_{s,d}(i, j)) = p(x_{s,d}(i, j)|H_1)/p(x_{s,d}(i, j)|H_0)$ and by expressing the shrinkage factor as $\hat{y}_{s,d}(i, j) = \gamma_{s,d}(i, j)/(1 + \gamma_{s,d}(i, j))w_{s,d}(i, j)$, where $\gamma_{s,d}(i, j) = \rho\xi(w_{s,d}(i, j))\eta(x_{s,d}(i, j))$ is the generalized likelihood ratio with $\rho = P(H_1)/P(H_0)$.

In this paper, we put the main idea of [1] into a fuzzy logic framework and develop a novel *FuzzyShrink* method. Namely, we also express the shrinkage factor for the wavelet coefficient $w_{s,d}(i, j)$ as a function of $w_{s,d}(i, j)$ and $x_{s,d}(i, j)$, but instead of estimating the likelihood ratios for these measurements, we impose on them *fuzzy membership functions*. Our shrinkage factor will also express how likely it is that a coefficient is a signal of interest, but we shall accomplish this by using the appropriate *fuzzy norms* and *co-norms* as opposed to the Bayesian formalism and probabilities.

2.1 Defining Membership Functions and a Fuzzy Rule

Our reasoning in defining the fuzzy shrinkage rule is the following. If both the neighbourhood around a given position (i, j) and the wavelet coefficient at this position itself ($w_{s,d}(i, j)$) contain mainly large (small) coefficients then we have enough indication that we have a signal of interest (noise). If the wavelet coefficient $w_{s,d}(i, j)$ is small but the neighbourhood around a given position (i, j) contains of mainly large coefficients then it is wise to give more importance to the neighbourhood instead wavelet coefficient $w_{s,d}(i, j)$ itself to judge if the value is a signal of interest or not. Otherwise we would give more importance to one single value (that does not correspond to the neighbourhood), which of course is less robust. In this situation (i.e. a small $w_{s,d}(i, j)$ but a large neighbourhood) we should conclude that the position (i, j) is a signal of interest, in spite of the fact that the coefficient is probably lower than the given threshold. This leads us to the Fuzzy Rule 1 introduced below, where the variable $x_{s,d}(i, j)$ represents the average of the wavelet coefficients in the $(2K + 1) \times (2K + 1)$ neighbourhood around a given position (i, j) . This variable indicates if the corresponding neighbourhood contains mainly large or small wavelet coefficients.

Fuzzy Rule 1. *The definition of the membership degrees in the fuzzy set **signal of interest** of the wavelet coefficient $w_{s,d}(i, j)$ with scale s and orientation d :*

IF $\left(|x_{s,d}(i, j)| \text{ is a large variable AND } |w_{s,d}(i, j)| \text{ is a large coefficient} \right)$

OR $|x_{s,d}(i, j)| \text{ is a large variable}$

THEN $w_{s,d}(i, j) \text{ is a signal of interest}$

Fuzzy rules are linguistic IF-THEN constructions that have the general form “IF A THEN B ”, where A and B are (collections of) propositions containing linguistic variables. A is called the premise or antecedent and B is the consequence of the rule. In Fuzzy Rule 1 we can distinguish two linguistic variables for the

consequent: (i) large wavelet coefficients $|w_{s,d}(i, j)|$ and (ii) large neighbourhood values $|x_{s,d}(i, j)|$. Both linguistic terms are modelled as fuzzy sets. A fuzzy set C [23] in a universe U is characterized by a $U - [0, 1]$ mapping μ_C , which associates with every element u in U a degree of membership $\mu_C(u)$ of u in the fuzzy set C . In the following, we will denote the degree of membership by $C(u)$.

The membership functions that are used to represent the two fuzzy sets of (i) large wavelet coefficient $|w_{s,d}(i, j)|$ and (ii) large neighbourhood value $|x_{s,d}(i, j)|$, are denoted as μ_w and μ_x , respectively. We use triangular membership functions shown in Fig. 1 (a) and (b).

From these figures we see that our method depends on three parameters. As in many image processing methods it is important that each filtering method is adapted to the noise situation (noise level). Therefore we have related all these parameters to the standard deviation of the noise. Good choices for the parameters are: $T_1 = \sigma$, $T_2 = 2\sigma$ and $T_3 = 2.9\sigma - 2.625$, with σ the standard deviation of the noise, which is estimated with the median estimator proposed by Donoho and Johnstone [25]. Those threshold values were obtained experimentally by optimising their performance on several test images with several noise levels.

The membership functions for the two fuzzy sets that are shown in Fig. 1 function as a kind of lookup-tables for the likelihood ratios of the probabilistic versions [1].

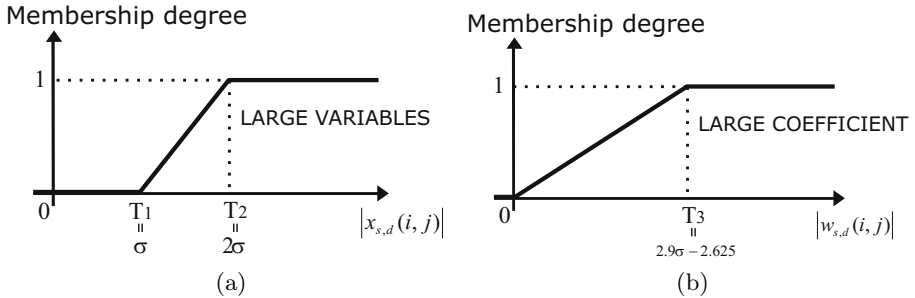


Fig. 1. (a) The membership function LARGE COEFFICIENT denoted as μ_w for the fuzzy set *large coefficient* and (b) The membership function LARGE VARIABLE denoted as μ_x for the fuzzy set *large variable*

In Fuzzy Rule 1 we can observe an intersection and a union of two fuzzy sets. The intersection $A \cap B$ of two fuzzy sets A and B is generally specified by a binary mapping D leading to: $(A \cap B)(y) = D(A(y), B(y))$. The union $A \cup B$ of two fuzzy sets A and B is specified by a binary mapping S leading to: $(A \cup B)(y) = S(A(y), B(y))$. In fuzzy logic, triangular norms (roughly the equivalent of AND operations) and triangular co-norms (roughly the equivalent of OR operations) are used to represent the intersection and the union of two fuzzy sets, respectively. Some well-known triangular norms together with their dual triangular co-norm are shown in Table 1.

Table 1. Some well-known triangular norms (T -norms) and triangular conorms (T -conorms)

T -norms	
minimum	$\min(x, y)$
algebraic product	$x \cdot y$
weak	$\begin{cases} \min(x, y) & \text{if } \max(x, y) = 1 \\ 0 & \text{otherwise} \end{cases}$
bounded sum	$\max(0, x + y - 1)$
T -conorms	
maximum	$\max(x, y)$
probabilistic sum	$x + y - x \cdot y$
strong	$\begin{cases} \max(x, y) & \text{if } \min(x, y) = 0 \\ 1 & \text{otherwise} \end{cases}$
bounded sum	$\min(1, \mu_A(x) + \mu_B(x))$

From all possible triangular norms the minimum norm is the largest and the weak norm (Table 1) is the smallest. From all possible triangular conorms the strong norm is the largest and the maximum norm (Table 1) is the smallest. We have chosen for a t -norm (with his dual conorm) which is situated between those two extremes, i.e. the product and the probabilistic sum, respectively. So the antecedent ($|x_{s,d}(i, j)|$ is **large variable** AND $|w_{s,d}(i, j)|$ is **large coefficient**) can be translated into the “truth” value: $\mu_x(|x_{s,d}(i, j)|) \cdot \mu_w(|w_{s,d}(i, j)|)$, where μ_x and μ_w are the membership functions for the fuzzy set *large variables* and *large coefficient*, respectively. In the next subsection we explain how to shrink the wavelet coefficients of a noisy image.

2.2 Output of the Method

The shrinkage rule of the proposed method for scale s , direction d and position (i, j) is calculated as follows:

$$\hat{y}_{s,d}(i, j) = \gamma(w_{s,d}(i, j), x_{s,d}(i, j)) \cdot w_{s,d}(i, j) \quad (3)$$

with $\hat{y}_{s,d}(i, j)$ the shrink output coefficient for scale s , direction d and position (i, j) and where $\gamma(w_{s,d}(i, j), x_{s,d}(i, j))$ is the degree of activation of Fuzzy Rule 1 for the wavelet coefficient $w_{s,d}(i, j)$. This value indicates the membership degree

in the fuzzy set *signal of interest* for the wavelet coefficient $w_{s,d}(i, j)$. If the membership degree has value 1, this means that the corresponding coefficient is a signal of interest certainly (and should not be changed), while a degree zero indicates that the coefficient is certainly not a signal of interest (and should be set equal to zero). A value between zero and one indicates that we do not know quite sure if this coefficient is a signal of interest or not. This means that the coefficient is a signal of interest only to a certain degree. The calculation of the value $\gamma(w_{s,d}(i, j), x_{s,d}(i, j))$ is illustrated in expression (4).

$$\gamma(w_{s,d}(i, j), x_{s,d}(i, j)) = \alpha + \mu_x(|x_{s,d}(i, j)|) - \alpha \cdot \mu_x(|x_{s,d}(i, j)|) \quad (4)$$

$$\text{with } \alpha = \mu_x(|x_{s,d}(i, j)|) \cdot \mu_w(|w_{s,d}(i, j)|)$$

Actually, the α of expression (4) can be seen as the fuzzy counterpart of generalized likelihood ratio used in the probabilistic version [1]. One can see that we used the product and probabilistic sum for the triangular norm and co-norm, respectively.

3 Experimental Results

In this section we present some experimental results. We compared our new fuzzy wavelet-based shrinkage method with (i) other well-known fuzzy filters and (ii) recently developed wavelet-based methods. More precisely we have:

- **FUZZY:** the GOA filter [26], FRINRM [27] (fuzzy randomly valued impulse noise reduction method), HAF [28] (histogram adaptive fuzzy), EIFCF [29] (extended iterative fuzzy control based filter), SFCF [29] (smoothing fuzzy control based filter), DWMAV [30] (decreasing weight fuzzy filter with moving average centre), AFSF [31] (the adaptive fuzzy switching filter), FSB [32,33] (fuzzy similarity filter) and AWFm [34] (adaptive weighted fuzzy mean).
- **WAVELET:** the bivariate wavelet shrinkage function proposed by Şendur [35], the feature-based wavelet shrinkage method proposed by Balster [36] and the probabilistic shrinkage function proposed by Pižurica [1].

We have used a redundant wavelet transform with the Haar wavelet and four resolution scales and a neighbourhood of size 9×9 ($K = 4$) for both the probabilistic version and the proposed one. As a measure of objective dissimilarity between a filtered image F and the original noise-free one O , we use the peak signal to noise ratio (PSNR).

In order to get a clear idea of the performance of all mentioned methods we have carried out experiments for three well known test images: ‘Lena’, ‘Peppers’ and ‘Barbara’, each of size 512×512 . The numerical results for the corrupted versions (for $\sigma = 5, 20, 30$ and 40) are shown in Table 2. From this Table we can make the following conclusions:

- The wavelet-based methods perform generally better than the state of the art fuzzy non-wavelet based methods for the additive noise type. Wavelet-based methods reduce the noise quite well for both low and high σ values, while the fuzzy-based methods only perform well for higher noise levels.
- The only fuzzy-based method that receives comparable results to the wavelet ones is the GOA filter. This filter even results in the best PSNR value for the Peppers images corrupted with $\sigma = 30$ and 40 additive Gaussian noise. But the GOA filter is developed only for a specific group of images like the Lena and the Peppers images. If an image contains regions with lots of fine details, texture or contours (like grass, hair etc.) then the GOA filter destroys such structures, which is confirmed by the low PSNR value for the Barbara image.
- Generally, the best numerical results were received by the proposed and the probabilistic shrinkage method. The proposed fuzzy shrinkage method performs quite similar as the probabilistic one.

The visual performance of the best numerical filters can be seen at <http://www.fuzzy.ugent.be/ACIVS05/paper161.pdf>, where we show (in Fig. 2) the denoised versions of the Barbara image corrupted with $\sigma = 40$ additive Gaussian noise. It is shown that the proposed and the probabilistic shrinkage method do not only yield the highest PSNR values (Table 2), but also the best visual results. The other wavelet-based methods reduce the noise well but introduce typical wavelet compression artefacts. From Fig. 2 (f) (in the longer paper version at the website) we see that the GOA filter destroys more images structures than the wavelet-based method, which results in a blurrier image. We can also conclude that the other state of the art fuzzy-based methods are not able to receive such good visual performances as the wavelet-based methods.

Previous experiments have clearly confirmed that the proposed method performs at least as well as the probabilistic method of [1]. In this paragraph we will illustrate that the proposed method, which can be viewed at <http://www.fuzzy.ugent.be/ACIVS06.html>, has a lower complexity than the probabilistic version. In Table 3 we have compared the execution time between those two methods for the noise reduction of one wavelet band of size 512×512 . The comparison is done by implementing both methods in the same programming language namely Java (not Matlab because Matlab uses many C-files so that the comparison would not be correct). The main difference of both methods is that the probabilistic method has to estimate the (image dependent) distributions first before the filtering can be started while the fuzzy shrinkage method can be applied directly. This fuzzy shrinkage method uses membership functions that are shown in Fig. 1, which functions as a kind of lookup-tables for the likelihood ratios of the probabilistic versions [1]. This explains while the proposed method is less complex. The execution time for the distribution estimation of [1] does not depend on the used neighbourhood size. Next even if we observe the execution time of the denoising methods only we see that the fuzzy shrinkage method is faster. This small difference is analysed in Table 4, where we have compared the amount

Table 2. PSNR results for the (512×512 -) *Lena*, *Peppers* and *Barbara* images corrupted with additive Gaussian noise with $\sigma = 5$, $\sigma = 20$, $\sigma = 30$ and $\sigma = 40$ and several fuzzy and wavelet based denoising methods

σ	Lena				Peppers				Barbara			
	5	20	30	40	5	20	30	40	5	20	30	40
Noisy	34.2	22.1	18.6	16.1	34.2	22.1	18.6	16.1	34.2	22.1	18.6	16.1
New	38.2	32.4	30.5	29.2	37.1	32.0	30.5	29.3	37.2	29.7	27.5	25.9
ProbShrink	38.3	32.3	30.4	29.2	37.1	32.0	30.5	29.3	37.2	29.4	27.1	25.5
BiShrink	37.4	31.2	29.3	28.1	35.7	31.0	29.3	28.1	36.2	28.2	26.1	24.8
Balster	37.2	31.5	29.8	28.5	34.4	31.7	30.1	28.9	35.8	27.6	25.3	24.0
GOA	36.4	31.2	29.5	28.3	35.6	31.7	30.0	28.6	33.9	25.8	24.2	23.5
FRINRM	34.9	26.2	23.8	21.3	34.3	25.4	22.2	20.4	34.2	23.7	21.4	20.2
HAF	33.7	29.5	26.9	24.8	33.1	28.8	26.2	24.0	25.3	24.4	23.3	21.1
EIFCF	33.6	29.3	27.2	25.5	33.8	29.5	27.3	25.6	25.5	24.6	23.7	22.8
SFCF	33.1	29.4	26.2	23.5	33.1	29.4	26.3	23.6	25.8	24.8	23.3	21.6
DWMAV	33.2	29.6	27.2	25.2	32.9	29.4	27.1	25.1	25.2	24.4	23.5	22.6
AFSF	34.5	27.6	25.0	23.0	34.4	27.6	24.9	22.9	26.0	23.9	22.5	21.2
FSB	33.8	28.8	25.5	23.1	33.7	28.9	25.7	23.3	25.2	23.9	22.6	21.2
AWFM	34.3	29.2	26.1	22.1	34.2	29.4	25.2	23.0	26.1	24.5	22.9	22.9

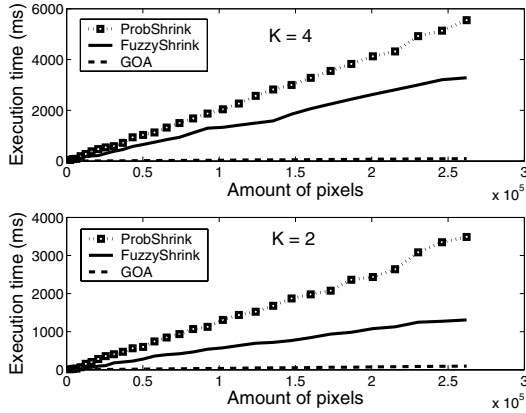


Fig. 2. Comparison of the mean execution time for the ProbShrink method of [1], the GOA filter of [26] and the proposed FuzzyShrink method with (upper) a window size $K = 4$ and (lower) a window size $K = 2$

of operations that have to be carried out to perform the denoising method for one wavelet band only. We observe that the amount of logical operations is very similar. But if we know that memory operations cause more time to be done than all other operations we see why the probabilistic method is slower.

In Fig. 2 we have illustrated the mean execution time of both methods (for a neighbourhood size of 5×5 and 9×9 (i.e. $K = 2$ and $K = 4$, respectively)) and the fuzzy non-wavelet based GOA filter [26]. We observe that the non-wavelet

Table 3. Comparison between the proposed fuzzy shrinkage method (FuzzyShrink) and the probabilistic shrinkage method (ProbShrink) in term of the execution time (ms) for the denoising method of a noisy wavelet band of size (512×512)

		Execution time in <i>ms</i>			
		$K = 1$	$K = 2$	$K = 3$	$K = 4$
FuzzyShrink	Total	58.5	108.0	179.3	273.5
ProbShrink	Denoising	63.4	110.6	195.8	282.0
	Distribution estimation	179.9	180.0	180.3	180.3
	Total	243.3	290.6	376.1	462.3

Table 4. Comparison between the proposed fuzzy shrinkage method (FuzzyShrink) and the probabilistic shrinkage method (ProbShrink) of the amount of operations necessary for the denoising methods of a noisy wavelet band of size $(N \times M)$ with $\eta = M \cdot N$ (exclusive the amount of operations necessary to calculate the distribution estimation)

		Execution time in <i>ms</i>				
		+	-	/	*	memory
FuzzyShrink		$(4 + (2K + 1)^2)\eta$	6η	3η	4η	$((2K + 1)^2)\eta$
ProbShrink		$(5 + (2K + 1)^2)\eta$	η	2η	5η	$((2K + 1)^2 + 3)\eta$

based method GOA performs much faster than the two wavelet based algorithms. The main two reasons for this difference are (i) in wavelet-based methods, the images have to be transformed into the wavelet domain and (ii) for both methods we have used a redundant wavelet transformation, so that the amount of data becomes larger. The second observation that can be made from Fig. 2 is that the proposed method is significantly faster than the probabilistic shrinkage method, which confirms that the proposed method is less complex than the probabilistic one.

4 Conclusion

In this paper an alternative wavelet based soft-computing method for the recently published probabilistic shrinkage method of Pižurica [1] for the reduction of additive Gaussian noise in digital images was proposed. Experimental results show that the proposed method receives the same noise reduction performance as the probabilistic one, which outperforms the current fuzzy-based algorithms and some recently published wavelet-based methods. Next we have shown that the proposed method clearly reduces the complexities of the probabilistic shrinkage method in terms of execution time. A future advantage of the method is the ability of incorporate more information (e.g. interscale and/or colour information) by adding other fuzzy rules to improve the noise reduction performance. Future work should be done on this promising issue.

Acknowledgment

This work was financially supported by the GOA-project 12.0515.03 of Ghent University.

References

1. Pizurica, A., Philips, W.: Estimating the probability of the presence of a signal of interest in multiresolution single- and multiband image denoising. *IEEE Transactions on Image Process.* **15**(3) (2006) 654-665
2. Resnikoff, H. L., Wells, R.O.: *Wavelet Analysis: The Scalable Structure of Information* Springer-Verlag (1998)
3. Donoho, D.: Denoising by soft-thresholding. *IEEE Transactions on Information Theory.* **41**(5) (1995) 613-627
4. Donoho, D., Johnstone, I.: Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association.* **90** (1995) 1200-1224
5. Chang, S., Yu, B., Vetterli, M.: Adaptive wavelet thresholding for image denoising and compression. *IEEE Transactions on Image Processing.* **9**(9) (2000) 1532-1546
6. Hansen, M., Yu, B.: Wavelet thresholding via mdl for natural images. *IEEE Transactions on Information Theory.* **46**(8) (2000) 1778-1788
7. Simoncelli, E., Adelson, E.: Noise removal via Bayesian wavelet coring. *Proceedings IEEE International Conference on Image Processing (ICIP'96)*. Lausanne, Switzerland (1996) 379-382
8. Moulin, P., Liu, J.: Analysis of multiresolution image denoising schemes using generalized gaussian and complexity priors. *IEEE Transactions on Information Theory.* **45**(4) (1999) 909-919
9. Pizurica, A., Philips, W., Lemahieu, I., Acheroy, M.: A joint inter- and intrascale statistical model for Bayesian wavelet based image denoising. *IEEE Transactions on Image Processing.* **11**(5) (2002) 545-557
10. Portilla, J., Strela, V., Wainwright, M., Simoncelli, E.: Image denoising using gaussian scale mixtures in the wavelet domain. *IEEE Transactions on Image Processing.* **12**(11) (2003) 1338-1351
11. Vidakovic, B.: Nonlinear wavelet shrinkage with bayes rules and bayes factors. *Journal of the American Statistical Association.* **93** (1998) 173-179
12. Chipman, H., Kolaczyk, E., McCulloch, R.: Adaptive Bayesian wavelet shrinkage. *Journal of the American Statistical Association.* **92** (1997) 1413-1421
13. Sendur, L., Selesnick, I.: Bivariate shrinkage functions for wavelet-based denoising exploiting interscale dependency. *IEEE Transactions on Signal Processing.* **50**(11) (2002) 2744-2756
14. Crouse, M., Nowak, R., Baranuik, R.: Wavelet-based statistical signal processing using hidden Markov models. *IEEE Transactions on Signal Processing.* **46**(4) (1998) 886-902
15. Romberg, J., Choi, H., Baranuik, R.: Bayesian tree-structured image modeling using wavelet-domain hidden markov models. *IEEE Transactions on Image Processing.* **10** (2001) 1056-1068
16. Malfait, M., Roose, D.: Wavelet-based image denoising using a markov random field a priori model. *IEEE Transactions on Image Processing.* **6**(4) (1997) 549-565
17. Jansen, M., Bultheel, A.: Empirical Bayes approach to improve wavelet thresholding for image noise reduction. *Journal of the American Statistical Association.* **96**(454) (2001) 629-639

18. Mihcak, M., Kozintsev, I., Ramchandran, K., Moulin, P.: Low complexity image denoising based on statistical modeling of wavelet coefficients. *IEEE Signal Processing Letters*. **6** (1999) 300-303
19. Fan, G., Xia, X.: Image denoising using local contextual hidden markov model in the wavelet domain. *IEEE Signal Processing Letters*. **8**(5) (2001) 125-128
20. Fan, G., Xia, X.: Improved hidden Markov models in the wavelet domain. *IEEE Transactions on Signal Processing*. **49** (2001) 115-120
21. Kerre, E.E.: *Fuzzy sets and approximate Reasoning*. Xian Jiaotong University Press (1998).
22. Tizhoosh, H.R.: *Fuzzy-Bildverarbeitung: Einfhrung in Theorie und Praxis*. Springer-Verlag (1997)
23. Zadeh, L.A.: Fuzzy Sets. *Information and Control* **8**(3) (1965) 338-353
24. Zadeh, L. A.: Fuzzy logic and its application to approximate reasoning. *Information Processing* **74** (1973) 591-594
25. Donoho, D.L., Johnstone, I.M.: Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** (1994) 425-455
26. Van De Ville, D., Nachttegael, M., Van der Weken, D., Kerre, E.E., Philips, W.: Noise reduction by fuzzy image filtering. *IEEE Transactions on Fuzzy Systems* **11**(4) (2003) 429-436
27. Schulte, S., De Witte, V., Nachttegael, M., Van der Weken, D., Kerre, E.E.: Fuzzy Random Impulse Noise Reduction Method. *Fuzzy Sets and Systems* (2006) (submitted)
28. Wang, J.H., Chiu, H.C.: An adaptive fuzzy filter for restoring highly corrupted images by histogram estimation. *Proceedings of the National Science Council - Part A* **23** (1999) 630-643
29. Farbiz, F., Menhaj, M.B., Motamedi, S.A.: Edge Preserving Image Filtering based on Fuzzy Logic. *Proceedings of the 6th EUFIT conference* (1998) 1417-1421
30. Kwan, H.K., Cai, Y.: Fuzzy filters for image filtering. *Proceedings of Circuits and Systems (MWSCAS-2002). The 2002 45th Midwest Symposium* (2002) III-672-5.
31. Xu, H., Zhu, G., Peng, H., Wang, D.: Adaptive fuzzy switching filter for images corrupted by impulse noise. *Pattern Recognition Letters* **25** (2004) 1657-1663
32. Tolt, G., Kalaykov, I.: Fuzzy-similarity-based Noise Cancellation for Real-time Image Processing. *Proceedings of the 10th IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)* **1** 15 -18
33. Tolt, G., Kalaykov, I.: Fuzzy-Similarity-Based Image Noise Cancellation. *Lecture Notes in Computer Science: Advances in Soft Computing (AFSS 2002)* **2275** (2002) 408-413
34. Kuo, Y.H., Lee, C.S., Chen, C.L.: High-stability AWFM filter for signal restoration and its hardware design. *Fuzzy Sets and Systems* **114**(2) (2000) 185-202
35. Şendur, L., Selesnick, I.W.: Bivariate Shrinkage Functions for Wavelet-based Image Denoising. *IEEE Transactions on Signal Processing* **50**(11) (2002) 2744-2756
36. Balster, E.J., Zheng, Y.F., Ewing, R.L.: Feature-based wavelet shrinkage algorithm for image denoising. *IEEE Transactions on Image Process.* **14**(3) (2005) 2024-2039

Mathematical Models for Restoration of Baroque Paintings

Pantaleón D. Romero and Vicente F. Candela

Department of Applied Maths. University of Valencia
Carrer Doctor Moliner 50, 46100, Burjassot (València)
pantaleon.romero@uv.es, vicente.candela@uv.es

Abstract. In this paper we adapt different techniques for image deconvolution, to the actual restoration of works of arts (mainly paintings and sculptures) from the baroque period. We use the special characteristics of these works in order to both restrict the strategies and benefit from those properties.

We propose an algorithm which presents good results in the pieces we have worked. Due to the diversity of the period and the amount of artists who made it possible, the algorithms are too general even in this context. This is a first approach to the problem, in which we have assumed very common and shared features for the works of art. The flexibility of the algorithm, and the freedom to choose some parameters make it possible to adapt the problem to the knowledge that restorators in charge may have about a particular work.

1 Introduction

Image restoration is a widely known topic in image processing, consisting of recover some deteriorated features of an image. But, before its use as a scientific subject, restoration is a term used in the artistic world meaning the task of cleaning and modifying a spoiled work of art in order to make it as similar as possible to the aspect it had when created.

Professionals of art restoration must have a multidisciplinary, both artistic and scientific training. Though it has been introduced recently, as in most other aspects, nowadays the help afforded by computer design packages of programs is very important. Restoration techniques are more and more complex, and it is required more and more exigency in the results. Thus, one cannot trust only in his or her own feelings in order to get the best possible results. Computer can never replace the human factor, but it can be used to provide support to the restorer, in order to recover possible lost elements or enhance details which can be important during the restoration process.

From a mathematical point of view, paintings may be consider as a particular class of images. Each period, and even each artist has its own characteristics, but there are some properties different from other images, such as photographs, medical images or other kind of images. We will comment some of these particularities. On the other side, deterioration of paintings is also due to some sort of physical processes, which we will analyze.

The goal of this paper is to obtain an algorithm to deconvolve artistic paintings. In particular, our study is restricted to Spanish baroque, and it has been applied to the altarpiece of the church of Saint Bartholomew in Bienservida. In our interest to make it as general as possible, we have suppressed some details that improve the results in this case but are very restrictive to be applied in other art works.

The structure of this paper is as follows. In the next section, we introduce the frame in which we raise our problem. In §3 some algorithms are introduced, and we establish a strategy in §4. Finally, in §5 results are presented, and we draw conclusions and prospectives.

2 The Problem of Deconvolving Paintings

Artistic restoration of paintings must consider two main tasks: one is inpainting (that is, to fill lost regions of the painting, often needing previous knowledge of the piece) and other one is cleaning. Cleaning is a process that includes enhancement and elimination of spurious details (dust particles, for instance). In this paper we address this last point. Practical application of mathematical models for inpainting deserve by themselves their own study.

The pieces we want to restore belong to the Spanish baroque. We are not going to get deeper into the artistic analysis, except for the special features we can deduce, having mathematical consequences, which are the following:

- In the artistic style we deal, the edges of the original pieces are not strictly well outlined, that is, some blur was present even when the work was painted (as a difference with photographs or drawings, or even other kind of paintings. Impressionism, or realism, for example, should need a different treatment).
- Weather factors (humidity, light, extreme temperatures, ...) and time spoil the works in a very deterministic way, and there is an almost absence of stochastic white noise, because dust, grease and other particles have created a uniform layer instead of a random distribution.

The first one allows us a broader definition of image enhancement, in the sense that we are worried about the location but not so much the enhancement of edges, in order to get *painting quality*. Without losing interest on edge enhancement, we will also be concerned, by the color and contrast enhancement. Besides, it is safe the use of the fast Fourier transform, because Gibbs phenomenon is controlled.

The second point makes us consider a very particular class of blur. Mathematically, the only noise we must be careful with is that produced by our own algorithms, and not by the image. Practically, it is a case of pure deconvolution without noise. The spoil is usually extended to large areas of the painting, and we may consider it global. Furthermore, the uniform appearance of the factors, provides the basis to consider that the blur has been produced by Gaussian-type convolution kernels.

The above remarks together with the usual image processing, are summarized in the following paragraphs.

We consider a gray scale image as a two dimensional function, $f(x, y)$, where (x, y) are the spatial coordinates, and the value is the normalized intensity of the light (from 0 for black, to 1 for white).

In a general frame, the image we have is the result of an original one $f_o(x, y)$, after being spoiled by a linear process:

$$f(x, y) = f_o(x, y) * k(x, y) + n(x, y)$$

k is called the kernel (or *psf*, point spread function) of the convolution, and it does not have to be spatially invariant (in fact, the spoil does not need to be linear. We assume so due to the special characteristics of painting restoration), but in practice we are going to consider it that way. k represents the deterministic part of the deterioration, and n , the noise, is the stochastic part.

As we said above, in our context, we may consider noise is not present, because the pictures have been cleaned by the professional restorers when we start the digital process. Then, we have:

$$f(x, y) = f_o(x, y) * k(x, y) = \int f_o(x - s, y - u)k(s, u)dsdu$$

The deconvolution problem consists of recovering the original function, f_o .

This is an ill conditioned problem ([7]). Linear convolution smoothes the function and, thus, irrecoverable information is lost (such as discontinuities, corners, ...). In the sense of Fourier transform,

$$\widehat{f} = \widehat{f}_o \widehat{k} \quad (1)$$

and \widehat{k} is a function decaying to zero in the large frequencies, making the above formula unstable if used to recover \widehat{f}_o .

Our problem is a blind deconvolution one. We do not only have to deconvolve the image, but we do not know what k is. However, in this case, we have some information about k . We know, for example, that k is almost Gaussian, in the sense that:

$$\widehat{k}(\xi, \eta) = e^{-\alpha(\xi^2 + \eta^2)^\beta} \quad (2)$$

for some $\alpha > 0$, $0 < \beta \leq 1$.

A kernel k satisfying (2) and defining fractionary powers

$$K^t f = \{\widehat{k}^t(\xi, \eta)\widehat{f}(\xi, \eta)\}^\vee, \quad 0 \leq t \leq 1$$

is said to be a class \mathcal{G} kernel ([2]) (as usual, \vee denotes the inverse Fourier transform). These kernels appear in generalized linear diffusion processes associated to the backward heat-like equation (with fractional powers of the Laplacian Δ):

$$\left. \begin{aligned} u_t &= - \sum_i \lambda_i (-\Delta)^{\beta_i} u & 0 < t \leq 1 \\ u(x, y, 1) &= f(x, y) \end{aligned} \right\} \quad (3)$$

where the coefficients λ_i are related to the exponent α of the kernel k , $\lambda_i = \alpha_i (4\pi^2)^{-\beta_i}$.

The above equation is ill-posed, because it is not reversible. In the next section it will be solved by introducing a regularizing term.

Paintings are a special kind of images, defined as the class \mathcal{N} , which their Fourier transforms verify, once normalized, that

1. $|\widehat{f}^*(\xi, \eta)|$ has isolated zeros in the frequency space, (ξ, η) .
2. $\log(|\widehat{f}^*(\xi, \eta)|)$ is decreasing along any ray $re^{i\theta}$, for r increasing, except for isolated singularities.

The normalization of \widehat{f} , is defined by:

$$\widehat{f}^*(\xi, \eta) = \frac{\widehat{f}(\xi, \eta)}{\widehat{f}(0, 0)}$$

due to the fact that $|\widehat{f}(\xi, \eta)| \leq \widehat{f}(0, 0)$, because of Hölder's inequality.

In the actual context, color is an important subject in order to get good results, and it must be taken account as a fundamental part of our study. In literature, color is obtained by decomposition of the image in three filters, each one with information of features of the chromatic components. Though in screening and printing it is advisable to decompose in primary colors (that is, RGB or CMYK), it is known that paintings must be repaired not because they lose color, but because ambiental factors affect their luminance, brightness or contrast. Thus, we get better results by decomposing the color in one luminance channel and two complementary ones. LUV and HSV filters are more adequate to our problem. In the experiments we have done, LUV has afforded better resolution than the others, due to the fact that there is less correlation between the luminance and the other channels. Enhancement filters produce an undesired effect on monochromatic decompositions (RGB or CMYK), in the sense that they increase the dominance of some colors above the others, modifying their equilibrium.

Now, we have the elements leading us to choose the right models to deconvolve the images. In the next paragraph, we introduce two models which will be combined in order to get acceptable results.

3 Deconvolution Models

Due to the characteristics of the kernel in our problem, we will start by estimate the parameters α and β in (2), by means of a nonlinear least-square approximation, in this way: as we consider there is practically no noise, from (1), after normalizing the image,

$$\log |\widehat{f}^*(\xi, \eta)| \approx -\alpha(\xi^2 + \eta^2)^\beta + \log |\widehat{f}_o^*(\xi, \eta)|$$

Of course, we do not know the values $|\widehat{f}_o^*|$. We will apply the so called APEX method ([2]), consisting of replacing $\log |\widehat{f}_o^*|$ by a fixed value A , determined by

the apex of $\log f^*(\xi, 0)$, because it is not restrictive to obtain the coefficients by fixing the ray $\eta = 0$,

$$\log |\widehat{f}^*(\xi, 0)| \approx -\alpha|\xi|^{2\beta} - A \quad (4)$$

By the secant method, we get the parameters α and β .

Once we have the kernel, we will use two direct deconvolution methods. As the blur is almost anisotropic, as we pointed out above, it is appropriate to consider one of them ([1]) based on the local linearization of the variational equation

$$-\Delta f_o + \lambda k * (k * f_o - f) = 0$$

λ is the Lagrange multiplier, which regularizes the equation in order to better conditioning. Large values of λ deconvolve the image, increasing the noise, while small λ smoothes it. In our case we shall choose λ as large as possible. By Fourier transform,

$$\widehat{f}_o(\xi, \eta) = \frac{\widehat{k}(\xi, \eta)\widehat{f}(\xi, \eta)}{\widehat{k}(\xi, \eta)^2 + \frac{4\pi^2(\xi^2 + \eta^2)}{\lambda}}$$

The other model we will use (Slow Evolution Constraint Backward, [2]) is the one coming from the generalized heat equation with fractionary powers (3). After regularizing, we look for solutions with a slow evolution, in the sense that

$$\|u(s) - u(0)\|_2$$

is bounded by a parameter, K depending on the regularization, the image and the estimation of the noise, in a given interval of points $[s^*, 1]$. This interval fixes the capability to recover $u(s)$ from the equation.

$$f^\dagger(\xi, \eta) = \frac{\widetilde{\widehat{k}}(\xi, \eta)\widehat{f}(\xi, \eta)}{|\widehat{k}(\xi, \eta)|^2 + K^{-2}|1 - \widehat{k}^s(\xi, \eta)|^2} \quad (5)$$

To obtain the solutions:

$$u^\dagger(x, y, t) = H^t f^\dagger(x, y)$$

Being a nonlinear method, this one, SECB, enhances edges in a finer scale than the first one, but, in order to be stable, we must choose between a loss of contrast as we approximate $u(s)$, for small, close to zero, s , due to a very large regularization parameter, or a very slight enhancement because small regularization parameters do not allow us to go back in time beyond an s close to 1.

4 Algorithm for Restoring Paintings

As we presented above, the local linearization and the SECB methods present different advantages and difficulties in order to deconvolve images. The algorithm

we introduce here tries to keep the advantages of both, while eliminating some of its problems. These are direct methods in the sense that we do not need any iterative process to reach the deconvolved image, a property that makes them appropriate for fast test and correction of the experimental parameters for the restorers in order to get a *good* deconvolution.

Local linearization limits the resolution till a certain scale, and beyond that the image cannot improve. On the other side, SECB may recover a finer resolution, but it must have a large regularization in order to keep stability. This regularization produces a loss of contrast. Our goal is to deconvolve fine scales without loss of contrast. To do this, we will follow the next steps, in what we may consider a hybrid method:

1. Detection of the parameters of the kernel by the algorithm in (4).
2. Deconvolution by the local linearization method.
3. Detection of the parameters of the kernel of the deconvolved image.
4. Deconvolution by the nonlinear method.

This algorithm may be consider as a multiscale process: we detect the kernel and deconvolve the image by a linear method in the coarse scales, and then we make the corresponding corrections in the finest scales by the nonlinear algorithm. This process allows us to improve the image with values of s close to 1 instead of close to 0, with a very small loss of contrast, because we do not need to recover the large frequencies, making it possible to enhance the edges by going back a very small interval of time. Thus, in this case, we may process images with a not so slow evolution.

As we process color images, we use the LUV filter, because it has the least correlation between luminance and colors. Thus, what we actually restore is the luminance without introducing spurious colors, which is an important point when restoring paintings.

Finally, we remark that both algorithms in our method may be performed via fast Fourier transforms. Thus, the proposed method is fast, and, due to both the regularity parameters and the features of the paintings (with practically no noise), it is stable. The idea of a linear method to enhance coarse scales and a nonlinear one to enhance the fine ones may be applied to other set of methods also (in fact, it is a sort of prediction-correction, or preconditioning of the image), but the characteristics of the algorithms we propose are adequate to the resolution we wish, and the computational features as speed, computational cost and stability we need in our context. In ([2]), the nonlinear method is proposed as the preconditioner of the image. Due to the previous application of the local linearization, SECB improves its performance and, in our case, gives well behaved results, without the need of more complex and expensive methods.

5 Results and Conclusions

The results we present in this section have been obtained from the altarpiece of *Iglesia de San Bartolomé* (Saint Bartholomew Church), in the town of Bienservida (Albacete, Spain), which was being restored while the elaboration of

this research. The photographs were taken at different stages of the restoration in the best conditions we could get. The altarpiece belongs to the *Obispado de Albacete* (Bishopship). The imagery is religious, as it happens with most of Spanish art of that era, and it is included in this work as a representative example of the application of the mathematical models, in its artistic, beyond its religious, context.

The paintings had a chemical treatment, cleaning, previous to our digital processing. This is the reason why the original pictures show a wide range of colours, though limited to the somewhat dark style of the Spanish Baroque. In these examples, we will show the enhancement obtained by our method, compared to the actual artistic restoration made by the professionals. In order to check the value of the preconditioning strategy, we will show the results obtained by the local linearization and by the SECB, in which their independent strengths and weaknesses are illustrated. We will see the improvement obtained by the combined, hybrid, strategy. The pictures we show were taken in the same technical conditions (light, distance, focus, ...). Though virtual restoration is not complete in our examples (we just did the deconvolution part), we show the artistic restored paintings in order to illustrate the task the professionals carry.

The goal of deconvolution is the enhancement of details in the painting which, are going to be reintegrated later in the process, interesting in order to establish sound boundary conditions for the mathematical models of segmentation to reconstruct wrought parts.

In the first example we show, corresponding to one of the paintings (*La Anunciación*), we see how both, linear and SECB, methods enhance edges, but, while the first one keeps contrast and luminance, the enhancement in the second one is sharper, in spite of its loss of brightness. This must be so, due to the large regularization parameter needed in order to go backward the equation. In our hybrid proposal we improve these methods, with a better enhancement and luminance quality. We also show the professional, *hand made* restoration. Though, as it was expected, this last was the best one, in some regions of the painting our restoration shows details which could not be observed in the original. It is remarkable that our method does not produce *ringing* effects nor additional noise in the processing, because we adjust boundary conditions and regularity is controlled. One of these regions may be examined in figure 2, in which we remark the lines written in the book, which are better outlined in the hybrid method than in any of the other ones, and even in the physical restoration by the professionals. Chemical color processing makes the restored image different from our convolved ones. Digital color processing is currently a work in progress in our research.

In figure 3, similar results are shown for another painting of the same set. The original picture is darker than the last one, but our method also works under these conditions. Some loss of luminance is obtained, but we can recover it after an independent treatment for contrast enhancement. Another of the advantages we present is the capability of our method to introduce different, mainly linear, processes independent of deconvolution without further deterioration of the image. This is an important requirement, because artistic restoration needs to



Fig. 1. (a) La Anunciación; (b) Local Linearization Method, $\lambda = 10^7$; (c) SECB, parameters: $s = 0.00001$, $K = 10^4$, $t = 0$; (d) Hybrid method parameters, $\lambda = 10^7$, $s = 0.1$, $K = 10^2$, $t = 0.9$; (e) After the artistic restoration

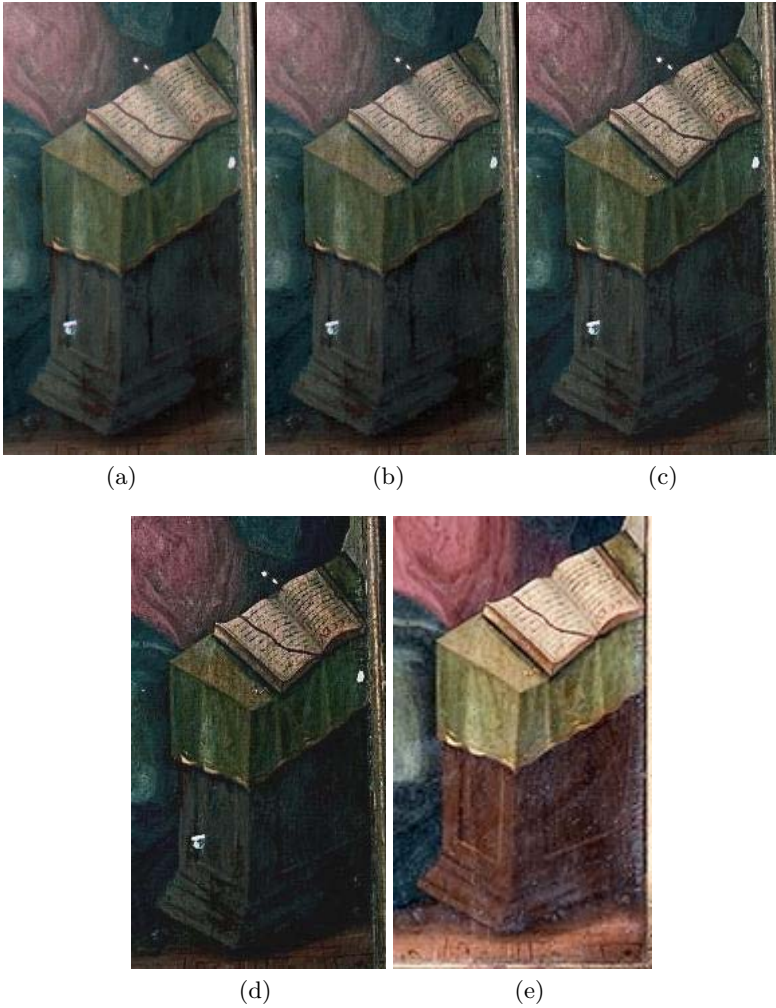


Fig. 2. (a) Section Desktop of La Anunciación; (b) Local Linearization Method; (c) SECB; (d) Hybrid method (e) After the artistic restoration

process different aspects of the artwork in a separate way. More sophisticated, and possibly better, methods lose some of their strengths when the image must be postprocessed (or preprocessed). We may observe that digital enhancement outlines edges with a greater accuracy than the actual restoration. In figure 3, it is observed how details in the clothes, mainly, are more remarkable in our method, as opposed as the actual artistic restoration. Chemical, and physical, color treatment, however, enlarges the quality aspect of the picture. Let us note that it was not the goal of this paper digital color processing, which is a different problem, that can be studied in further works.

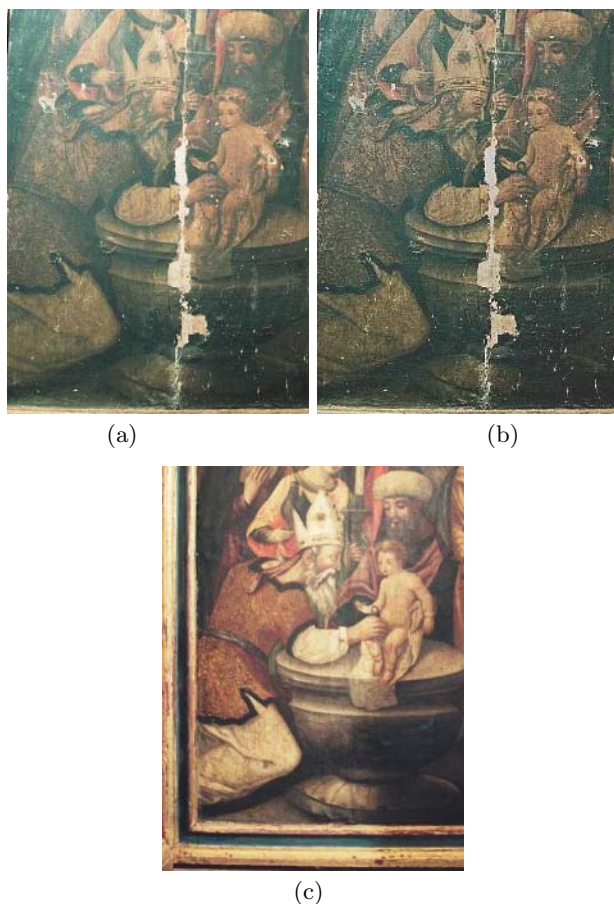


Fig. 3. (a) Paint of Circumcision; (b) Hybrid method parameters used $\lambda = 10^7$, $s = 0.1$, $t = 0.93$, $K = 10^3$; (c) After the artistic restoration

As a conclusion, we have devised a numerical algorithm for image deconvolution which is both reliable and robust for a large class of paintings. It allows different processes independent from deconvolution without loss of accuracy and it is fast enough to be performed with a low computational cost and time. It is possible the adaptation to factual requirements such as increase/decrease of luminance and color properties. Last, but not least, the results show acceptable human vision quality, and they may help restorers in their recovery tasks.

Acknowledgements

The authors thank J. Ángel Navarro, priest of *Iglesia de San Bartolomé* in Bimenservida, for his support in the documentation, and the permissions to obtain

the pictures of the altarpiece, as well as his disposal to help in logistic and personal aspects.

Thanks are also due to *Grupo Ábside*, restorers of the piece, in particular to their director, Guadalupe, whose professional advice and explanation of their job were useful throughout the research.

Comments and suggestions by J.V. Arnau Córdoba helped to improve the mathematical background of this paper.

This research has been supported by DGCYT project MTM2005-07708.

References

1. Candela V., Marquina A., Serna S., *A Local Spectral Inversion of a Linearized TV Model for Denoising an Deblurring*, IEEE Transactions on Image Processing, 2003, 12,7, pp.808-816
2. Carasso A., *Direct Blind Deconvolution*, SIAM J. Numer. Anal., 2001, 61, pp. 1980-2007
3. Osher S. J., Sethian J. A., Fronts propagation with curvature dependent speed: Algorithms based on Hamilton-Jacobi formulations, *Journal of computational physics*, 79 (1988), 12-49.
4. Perona P., Malik J., Scale-space and edge detection using anisotropic diffusion, *IEEE Transactions on pattern analysis and machine intelligence*, 12 (1990)
5. L.Rudin and S.Osher, *Total variation based image restoration with free local constraints*, Proc.IEEE Internat. Conf.Imag.Proc.,(1994), pp. 31-35.
6. A.Marquina and S.Osher, *Explicit algorithms for a new time dependent model based on level set motion for nonlinear deblurring and noise removal*, SIAM J. Sci. Comput., 22 (2000), pp. 387-405.
7. A. Tichonov and V. Arsenin, *Solution of ill-posed problems*, Wiley, New York, 1977.
8. F. Catté, P.L. Lions, J.M. Morel and T. Coll, *Image selective smoothing and edge detection by nonlinear diffusion*, SIAM J. Numer. Anal., 29 (1992), pp. 182-193
9. S.A. Martucci, *Symmetric convolution and the discrete sine and cosine transforms*, IEEE Trans. Signal Processing, 42 (1994), pp. 1038-1051
10. S. Bergeon, *Colour et restauration*, Techne, 4, 17-28, 1976
11. G. Wyszecky and W.S. Stiles, *Color Science*. John Wiley & Sons, New York, 372 (1982)

Motion Blur Concealment of Digital Video Using Invariant Features

Ville Ojansivu and Janne Heikkilä

Machine Vision Group, Department of Electrical and Information Engineering
P.O. Box 4500, FIN-90014 University of Oulu, Finland
{vpo, jth}@ee.oulu.fi

Abstract. This paper deals with concealment of motion blur in image sequences. The approach is different from traditional methods, which attempt to deblur the image. Our approach utilizes the information in consecutive frames, replacing blurred areas of the images with corresponding sharp areas from the previous frames. Blurred but otherwise unchanged areas of the images are recognized using blur invariant features. A statistical approach for calculating the weights for the blur invariant features in frequency and spatial domains is also proposed, and compared to the unweighted invariants in an ideal setting. Finally, the performance of the method is tested using a real blurred image sequence. The results support the use of our approach with the weighting scheme.

1 Introduction

In many applications, the quality of video is degraded by blurring. One important class of blurring is caused by camera motion during exposure. A typical situation which generates such blur is recording using a handheld camera. The blur becomes even more noticeable after video stabilization as the random motion no longer hides the blur. In this paper, we consider the concealment of this kind of global blur in stabilized video sequences.

Image deblurring is a difficult problem. Traditional solutions include the estimation of the point spread function (PSF) of the blur and deconvolution of the image using that PSF. When the PSF is known, the latter ill-posed problem can be solved using approaches which use regularization [1]. In practice, the PSF is unknown and very hard to estimate accurately. In this case, blind image restoration algorithms are used. Such algorithms are presented in [2].

In the case of an image sequence, it is natural to utilize multiple frames for restoration. This is the case in multichannel or multi-frame models which use sequential frames which are degraded differently. A survey of these approaches, as well as a method that does not need any information on the PSF or perfect registration of the frames are given in [3].

Our approach is also a multi-frame method in the sense that it utilizes the information in multiple frames. However, the approach is relatively simple and many obstacles of the above mentioned methods are avoided. The basic idea is to replace the scene which is detected to be blurred using the same areas from

a registered previous sharp frame. If the image is heavily blurred it is difficult to classify the image regions into unchanged scene and moving objects. We have divided the image into blocks and used blur invariant features invented by Flusser et al. [4] to classify the blocks as changed or unchanged for detection of motion. A partly similar deblurring method was presented in [5] but it uses the sum of absolute differences (SAD) of images as a classification criterion. Blur invariant features seem to work much better than SAD which is error prone in the case of heavy blur.

Previously, no weighting scheme of the Flusser's blur invariants has been presented. Our statistical approach weights the invariants according to the image noise characteristics. The weights are derived both for spatial and frequency domain blur invariants. The experiments confirmed that the invariants perform much better when the noise is taken into account. However, the border effect of the blur still remains a problem.

The rest of this paper is organized as follows. In Sect. 2, after an introduction of the blur invariant features, the method for estimation of the weight factors is proposed. Then in Sect. 3, the framework for concealment of motion blur in image sequences is presented. In Sect. 4, the performance of the statistical weighting of the blur invariants is compared to the unweighted use of the invariants [4] in an ideal setting. Finally, the results with our motion blur concealment framework are presented and compared to a reference method in a practical case.

2 Blur Invariants for Noisy Images

In this section, the invariants in the spatial and frequency domain are given and then the noise covariances in both cases are derived.

2.1 Blur Invariant Features

The blur invariant features of Flusser et al. [4] are invariant to blur that has centrally symmetric PSF, as is the case with ideal linear motion blur. If $f(x, y)$ is an image function, the blur invariants based on image moments in the spatial domain are the following:

$$I_S(p, q) = m_{pq} - \frac{1}{m_{00}} \sum_{n=0}^p \sum_{\substack{m=0 \\ 0 < n+m < p+q}}^q \binom{p}{n} \binom{q}{m} I_S(p-n, q-m) \cdot m_{nm}, \quad (1)$$

when order $r = p + q$ is even and $I_S(p, q) = 0$ otherwise, and where

$$m_{pq} = \sum_{i=1}^N \sum_{j=1}^N i^p j^q f(i, j) \quad (2)$$

is the approximation of the image moment of order $p+q$ and $N \times N$ is the image size.

Let $F(u, v)$ be the 2-D Fourier transform of the image $f(x, y)$. Then the blur invariants in the frequency domain are given by

$$I_F(u, v) = \tan(\text{ph}F(u, v)) = \frac{\text{Im } F(u, v)}{\text{Re } F(u, v)} = \frac{F_{\text{im}}(u, v)}{F_{\text{re}}(u, v)}. \quad (3)$$

The space and frequency domain invariants are equivalent from the theoretical point of view, as shown in [4], but may differ in numerical behavior, noise robustness etc. In a practical case, these invariant features are not fully invariant because of the boundary effect of convolution by the blur PSF. This is due to the fact that information flows across the borders of the observed image, as is shown mathematically in [6] for the invariants in the spatial domain. To be fully invariant, the whole convolution result would have to be known, which is impossible in practice.

2.2 Estimation of the Noise of the Blur Invariants

Image noise degrades the classification results obtained using the invariants. It is known that in the spatial domain case, the higher order moments suffer more from noise [7]. Previously any noise modeling has not been incorporated into these blur invariants [4,8]. We have derived a statistical approach which calculates weights for the invariants according to their estimated noise.

Let's assume the following model for an observed image:

$$\hat{f}(x, y) = f(x, y) + n(x, y), \quad (4)$$

where $f(x, y)$ is the original, possibly blurred image and $n(x, y)$ is zero-mean independent and identically distributed noise with variance σ^2 .

We will consider first the spatial invariants. If the image moments which are used to build invariants are calculated according to (2) for the observed images (4), the noisy moments are

$$\hat{m}_{pq} = \sum_{i=1}^N \sum_{j=1}^N i^p j^q \{f(i, j) + n(i, j)\} = m_{pq} + \sum_{i=1}^N \sum_{j=1}^N i^p j^q n(i, j). \quad (5)$$

It is easy to show that the covariances of the noisy moments are

$$\sigma_{pq,rs} = \sum_{i=1}^N \sum_{j=1}^N i^{p+r} j^{q+s} \sigma^2, \quad (6)$$

where p, q and r, s are the orders of the two moments for which the noise covariance is being calculated.

The covariance matrix for the spatial invariants cannot be calculated directly as the equations for the invariants are non-linear. However, the covariance matrix can still be approximated using linearization according to the equation

$$\mathbf{C}_r \approx \mathbf{J} \cdot \mathbf{C}_m \cdot \mathbf{J}^T, \quad (7)$$

where \mathbf{C}_m is the $N_m \times N_m$ covariance matrix build using covariances of equation (6) and \mathbf{J} is $N_r \times N_m$ Jacobian matrix containing partial derivatives

$$\frac{\partial \mathbf{I}_{S_i}}{\partial \mathbf{m}_j}, i = 1, \dots, N_r \text{ and } j = 1, \dots, N_m. \quad (8)$$

where N_r is the number of invariants up to order r and N_m is the number of moments needed in calculation of those invariants.

The $N_r \times N_r$ noise covariance matrix \mathbf{C}_r is used to weight the differences of the invariants by calculating the Mahalanobis distance between invariants $I_S(p, q)^{(f)}$ and $I_S(p, q)^{(g)}$ of images $\hat{f}(x, y)$ and $\hat{g}(x, y)$, namely

$$distance = \mathbf{D} \cdot \mathbf{C}_r^{-1} \cdot \mathbf{D}^T, \quad (9)$$

where

$$\mathbf{D} = \mathbf{I}_S^{(f)} - \mathbf{I}_S^{(g)}. \quad (10)$$

The use of the covariance matrix \mathbf{C}_r effectively weights each invariant according to its estimated signal-to-noise ratio. Notice that σ^2 is not needed if only relative distances are considered.

For the frequency domain invariants, the derivation of the noise covariance matrix is quite similar. The real and imaginary parts $F_{im}(u, v)$ and $F_{re}(u, v)$ of the Fourier domain invariants (3) are obtained as real and imaginary parts of the equation

$$\hat{F}(u, v) = \sum_{x=1}^N \sum_{y=1}^N \{f(x, y) + n(x, y)\} e^{-2\pi j(ux+vy)/N} \quad (11)$$

$$= F(u, v) + \sum_{x=1}^N \sum_{y=1}^N n(x, y) e^{-2\pi j(ux+vy)/N}, \quad (12)$$

which describes the discrete Fourier transform of a noisy image (4). The corresponding noise covariance between image frequencies (u_1, v_1) and (u_2, v_2) becomes

$$\sigma_{u_1 v_1, u_2 v_2} = \sum_{x=1}^N \sum_{y=1}^N e^{-2\pi j(u_1 x + v_1 y)} e^{-2\pi j(u_2 x + v_2 y)} \sigma^2. \quad (13)$$

It can be shown that the covariance is zero if (u_1, v_1) and (u_2, v_2) are different. Hence, only variances need to be calculated, leading to the simplification of (13), namely

$$\sigma_{uv} = \sum_{x=1}^N \sum_{y=1}^N e^{-2\pi j(ux+vy)} \sigma^2. \quad (14)$$

The resulting $2N_r \times 2N_r$ diagonal noise covariance matrix \mathbf{C}_m for $F_{im}(u, v)$ and $F_{re}(u, v)$ contains the diagonal values $\{Im(\sigma_{(uv)_1}), Re(\sigma_{(uv)_1}), \dots, Im(\sigma_{(uv)_{N_r}}), Re(\sigma_{(uv)_{N_r}})\}$ from equation (14). N_r is the number of Fourier domain invariants used.

The noise covariance matrix \mathbf{C}_r for the invariants (3) is calculated using linearization similar to (7) in the spatial domain. In this case, \mathbf{J} is a $N_r \times 2N_r$ matrix containing the partial derivatives of the invariants $I_F(u, v)$ with respect to $F_{im}(u, v)$ and $F_{re}(u, v)$. The distance between images is calculated similarly to the spatial case using equation (9).

3 Framework for Motion Blur Concealment

Typically not all the images in a sequence recorded with a handheld camera are severely blurred. Our idea is that after the blurred images are detected, the blurred but otherwise unchanged scene in these images is replaced using the same scene from previous sharp frames which are first registered. To replace only the unchanged parts of the scene, each image block is classified as changed (e.g. moving objects) or unchanged using the blur invariant features presented in Sect. 2 for motion detection.

Our framework for motion blur concealment is presented in Fig. 1(a). As can be seen, it consists of four steps preceded by stabilization, which is not considered here. However, the motion parameters that are needed for registration can be taken directly from the stabilization step. Otherwise they need to be estimated separately.

Detection of blurred frames is performed using gradient information. The sum of absolute gradients was approximated for each image $f_t(x, y)$ in a sequence to get the measure for its sharpness, namely

$$s_t = \sum_{x,y} \{ |d_x * f_t(x, y)| + |d_y * f_t(x, y)| \}, \quad (15)$$

where d_x and d_y are derivative filters along the x - and y -directions.

Measure s_t does not give an absolute evaluation of the image sharpness but if the scene is relatively unvarying this measure can be compared to the same measure for surrounding frames. A frame $f_t(x, y)$ is classified as blurred if

$$s_t < T \cdot \max\{s_{t-1}, \dots, s_{t-K}\}, \quad (16)$$

where T is a threshold and K is the number of previous images used as a reference. This approach is robust if the scene does not vary too quickly and some of the previous K frames are sharp. In our tests $T = 0.75$ and $K = 5$.

The classification of the scene into changed and unchanged blocks using blur invariant features is the main contribution of this research. The principle is as follows: The last image classified as sharp in the blur detection step is kept until the next one is encountered. This image is registered with the subsequent blurred frames. After this, the sharp and blurred frames are divided into blocks of size

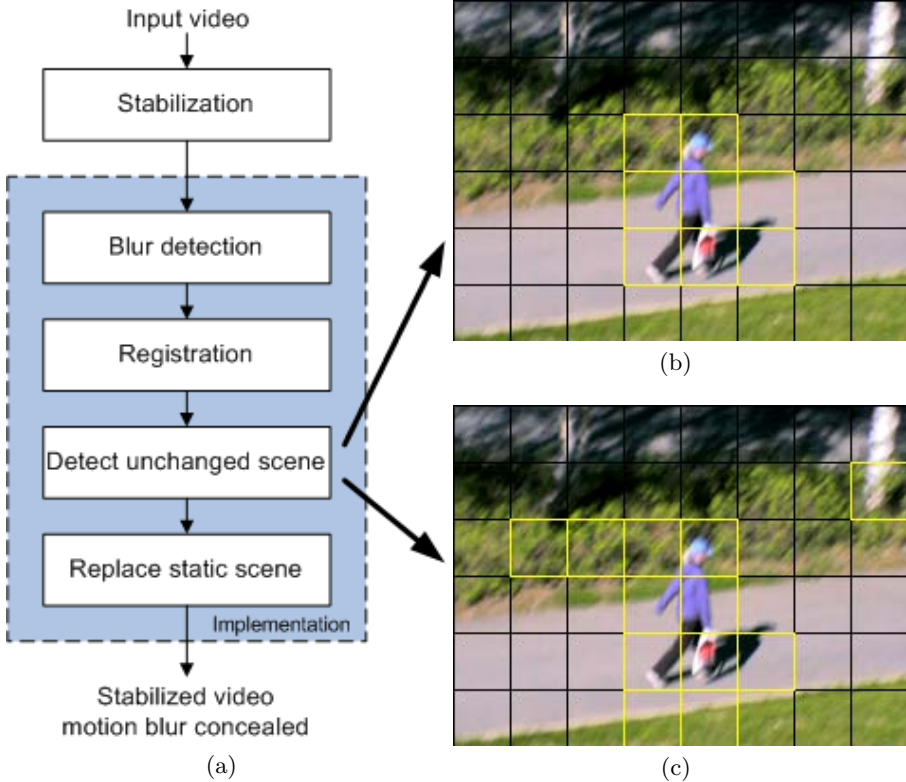


Fig. 1. (a) Framework for motion blur concealment. Classification into changed and unchanged scene in (b) using blur invariants and in (c) using SAD.

$B \times B$ as in Fig. 1(b) and 1(c), and the distance (10) using the blur invariant features is calculated between each sharp vs. blurred block pair. If this distance between corresponding blocks in sharp and blurred frames deviates enough, the blocks are considered to be different. In the last step, the unchanged blocks in blurred frame are replaced using the blocks from the sharp frame. We have used either spatial or frequency domain invariants, as discussed in Sect. 4. Fig. 1(b) shows an example of the classification of the blocks of a blurred image. The blocks with lighter borders are classified to be changed i.e. they are different in this and the preceding sharp image. Fig. 1(c) shows the classification result if SAD measure, which is not invariant to blur, is used.

The unchanged scene blocks are replaced using a method resembling that used in overlapped block motion estimation (OBME) [9] i.e. the overlapping blocks are larger than $B \times B$ and are windowed so that the total weight for each pixel is one. This makes the transition at the borders smooth. Fig. 2(a) illustrates a typical blurred input frame and Fig. 2(b) the corresponding output frame generated using our method to conceal the blur. As can be seen, all other blocks



Fig. 2. (a) Blurred input frame. (b) Motion blur concealed output frame.

except those containing the person are replaced and the transition to the blocks containing the person is smooth.

4 Experiments

The evaluation of deblurring methods in general is difficult as the correct unblurred video sequence is not known in a practical case for comparison. In our case, it is quite clear that the final quality of the video depends on the robustness of finding the unchanged blocks from each blurred frame. For this purpose, we have tested the classification performance of the unweighted and weighted blur invariant features in two experiments using an ideal and real setting.

In the experiments, we used a QVGA resolution test sequence of length 200 frames. The sequence was recorded using a vibrating handheld camera resulting in blurred frames like that shown in Fig. 2(a). The sequence was stabilized before the tests. The images of the sequence were divided into 6×8 blocks of size 32×32 pixels discarding the excess edge pixels. In both tests, we used 12 invariants in either the spatial or Fourier domain. This corresponds to spatial invariants of orders $r = 1, 3$ and 5 . In the Fourier domain, this yields to the invariants $I_F(u, v)$ for which $\sqrt{u^2 + v^2} \leq \sqrt{8}$, where u and v may also be negative, and when invariants corresponding to zero frequency and redundant mirror frequencies are discarded. In the unweighted case, the frequency domain invariants are taken as such and the spatial domain invariants are normalized by $(N/2)^{p+q}$, as proposed in [4], because the numerical range of the higher order spatial invariants is larger.

In the first experiment, we compared the classification results of the unweighted and weighted invariants in a theoretical setting in which we used artificially created blur. Before blurring, the image blocks of the test sequence were padded with zeros to cancel the border effect. In this way, the only factor deteriorating the results was the noise that was added after blurring. As the weighting of the invariants is based on the estimated noise in them, this was an ideal experiment to demonstrate the performance difference of the weighted and unweighted invariants.

Two classes of block pairs were created, based on the image blocks of the gray scale version of the test sequence: “unchanged” i.e. original vs. blurred blocks and “changed” i.e. original vs. shifted and blurred blocks as shown in Fig. 3. The class original vs. shifted blocks mimics the situation of a changing scene in the real video. The shift length was one pixel and the blur length 10 pixels. The noise standard deviation was either $\sigma = 2$ or $\sigma = 5$. Notice that the length of the blur does not have much effect in this ideal case.

In Fig. 4(a) and Fig. 4(b) the classification results are presented as receiver operating characteristics (ROC) curves for spatial and frequency domain invariants, respectively. The vertical axes show the probability of being correctly classified as “unchanged” and horizontal axis the probability of being incorrectly classified as “unchanged”. The results were calculated using either weighted or unweighted blur invariants and noise standard deviations $\sigma = 2$ or $\sigma = 5$. It can be seen that the weighting improves the results significantly. Particularly the frequency domain invariants seem to be performing unsatisfactorily for our purpose without weighting. Also in the case of spatial domain invariants the improvement in the classification accuracy is clear.

In the second experiment, the real test sequence containing only the original blur was used. 67 frames from the total of 200 were detected as blurred. The blocks of the test sequence were labeled subjectively into “changed” and “unchanged”. The classification result obtained using blur invariant features was compared to this ground truth classification to obtain similar ROC curves to those in the first experiment. True positive, i.e. correctly classified as unchanged scene, versus false positive, i.e. incorrectly classified as unchanged scene probabilities for different decision boundaries are shown in Fig. 5. Notice that for finer detail only the upper left quadrant of the full ROC diagram is shown. The curves represent the results using spatial or frequency domain invariants either weighted or unweighted. The classification results were compared to a simple change detection method based on the SAD of the image blocks. We calculated the blur invariants and SAD of blocks separately for each RGB-channel, and used their mean value as the final distance. Furthermore, the classification results were calculated only for the blurred frames of the sequence, as this is the interesting situation.

To get acceptable video quality the false positive rate has to be quite low. Otherwise the blocks containing moving objects may be replaced, contaminating the image content of the sequence. So one should concentrate on the results where the false positive rate is relatively low. First of all the results show the positive effect of weighting the invariants. The improvement of the true positive rate at a given false positive rate using weighted invariants in the spatial or frequency domain is not as large as in the ideal case, probably because of the boundary effect which now causes the results to deteriorate. The improvement is still significant in the case of spatial invariants, and again very large in the case of frequency domain invariants. Clearly, without weighting, the frequency domain invariants could not be used for our purpose.

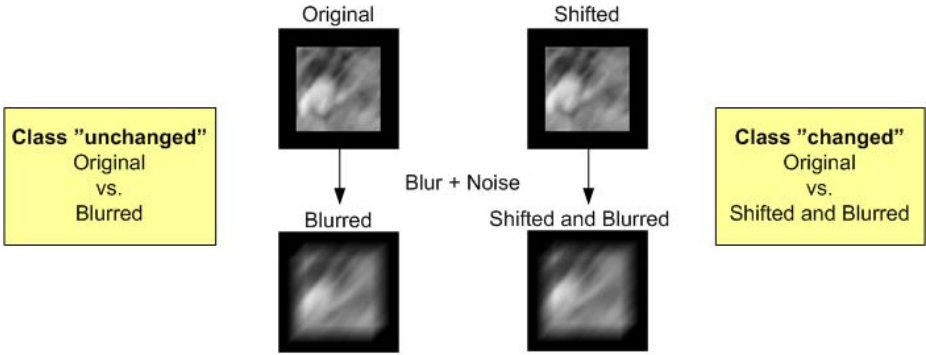


Fig. 3. Example of the blocks used for the two test classes

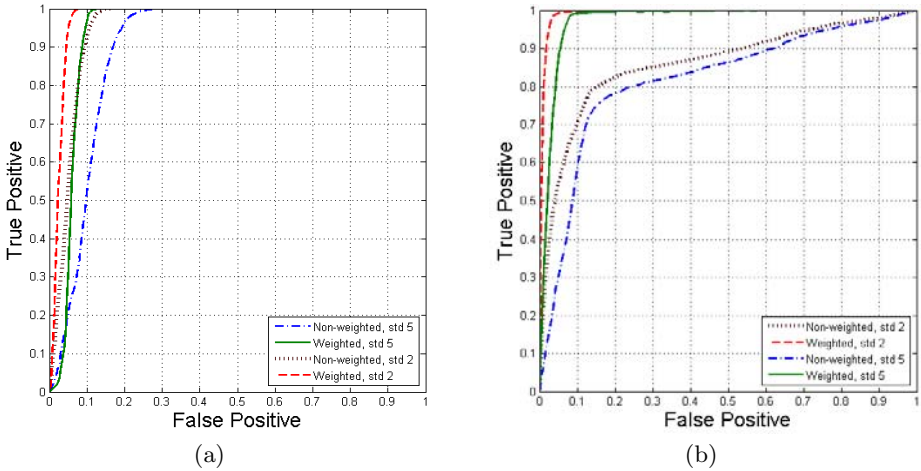


Fig. 4. (a) ROC curves for the spatial domain invariants. (b) ROC curves for the frequency domain invariants.

Except in the unweighted frequency domain case, all the invariant approaches performed better than SAD. This can be explained by the fact that SAD is not invariant to blur, and it separates blurred and unblurred blocks. According to tests in [4] the error of the boundary effect is significant when the ratio $blurlength/B$ is greater than 0.15. This corresponds to a blurring length of 4.8 pixels when $B = 32$. It was noticed also in our experiments that the blur invariants fail when the blur becomes too large.

The performance of the weighted frequency domain invariants seemed to be lower compared to the spatial case. On the other hand, the frequency domain invariants are much faster to calculate using FFT. Also the calculation of the weights is faster as the matrix \mathbf{C}_m is diagonal and the matrix \mathbf{J} block diagonal. For this reason we tried to improve the results. It was noticed that by using

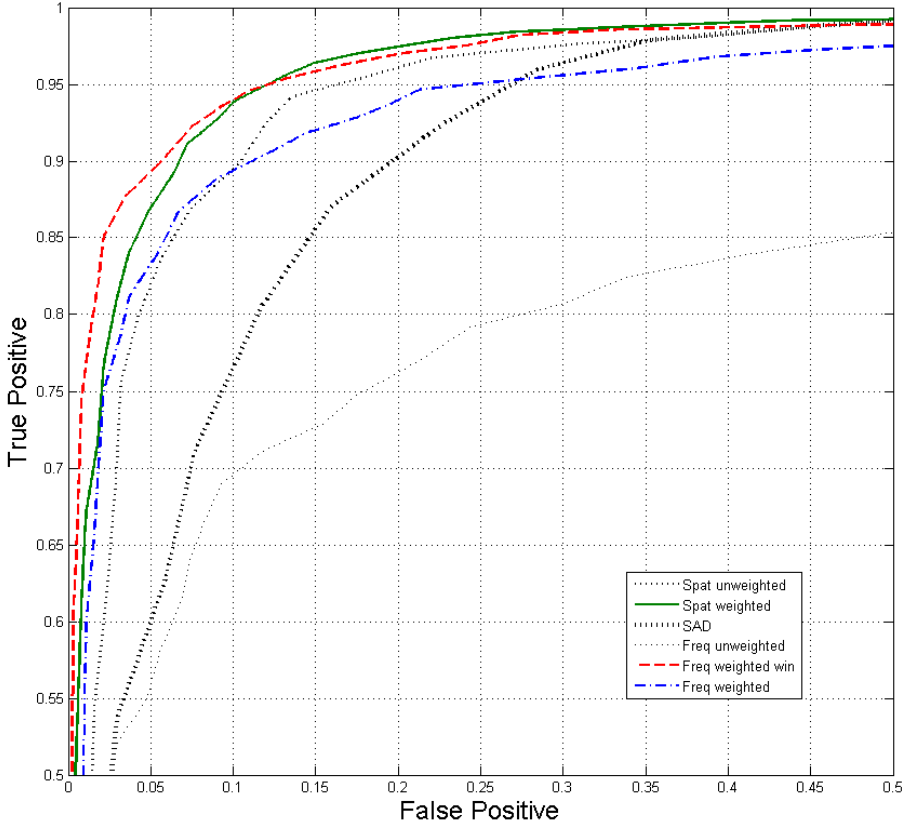


Fig. 5. ROC curves showing the classification performance of the invariants for the real world test sequence

slightly larger and overlapping image blocks and by windowing those blocks before calculation of the invariants, the results were improved as shown in Fig. 5 by curve “freq weighted win”. Here, the block size was 48×48 , overlapping 8 pixels, and the window was $W = \mathbf{k}^T \cdot \mathbf{k}$, where \mathbf{k} is a 1-D Kaiser filter, with length 48 and parameter $\beta = 2$. In the case of spatial domain invariants, the windowing did not help noticeable. When the false positive rate is less than 0.1 the weighted frequency domain invariant method using windowing seems to be the best alternative.

5 Conclusions

The method proposed in this paper can be used to conceal the motion blur of the unchanging scene of the video by replacing the blurred areas from the previous frames. The blur invariant features, which are used for recognition of

this unchanged scene, perform significantly better when our weighting scheme is used. It is noticeable that the frequency domain invariants, which are much faster to calculate, would be unsuitable for our purpose without weighting. When windowing is combined with weighting of the frequency domain invariants they become the best alternative. The invariants outperform also the SAD measure used as a reference. In practice, better classification of the blocks results in more complete concealment of the blur.

A downside of our method is that it cannot deblur the moving objects and other changing parts of the video. However, inclusion of a desirable deblurring method for moving objects is possible. The use of our method for as large region as possible is advantageous, as nuisances like ringing and noise amplification of the traditional deblurring methods are avoided.

Future improvements to the method might include calculation of the invariants for hierarchical and/or overlapping blocks. The invariants would first be calculated for larger blocks and blocks that seem to be changing would be investigated in greater detail. This is likely to reduce the border effect in the results and will make the calculation faster.

References

1. Banham, M.R., Katsaggelos, A.K.: Digital image restoration. *IEEE Signal Processing Magazine* **14** (1997) 24–41
2. Kundur, D., Hatzinakos, D.: Blind image deconvolution. *IEEE Signal Processing Magazine* **13** (1996) 43–64
3. Sroubek, F., Flusser, J.: Multichannel blind deconvolution of spatially misaligned images. *IEEE Transactions on Image Processing* **14** (2005) 874–883
4. Flusser, J., Suk, T.: Degraded image analysis: An invariant approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **20** (1998) 590–603
5. Matsushita, Y., Ofek, E., Tang, X., Shum, H.Y.: Full-frame video stabilization. In: *IEEE International Conference on Computer Vision and Pattern Recognition*, San Diego, CA, USA (2005) 50–57
6. Candocia, F.M.: Moment relations and blur invariant conditions for finite-extent signals in one, two and N-dimensions. *Pattern Recognition Letters* **25** (2004) 437–447
7. Pawlak, M.: On the reconstruction aspects of moment descriptors. *IEEE Transactions on Information Theory* **38** (1992) 1698–1708
8. Flusser, J., Boldys, J., Zitová, B.: Moment forms invariant to rotation and blur in arbitrary number of dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25** (2003) 234–246
9. Wang, Y., Ostermann, J., Zhang, Y.Q.: Video processing and communications. Prentice-Hall signal processing series. Prentice Hall, Upper Saddle River, New Jersey (2002)

Hybrid Sigma Filter for Processing Images Corrupted by Multiplicative Noise

Nikolay Ponomarenko¹, Vladimir Lukin¹,
Karen Egiazarian², Jaakko Astola², Benoit Vozel³, and Kacem Chehdi³

¹ National Aerospace University, Dept of Transmitters, Receivers and Signal Processing,
17 Chkalova St, 61070 Kharkov, Ukraine
uagames@mail.ru, lukin@xai.kharkov.ua

² Tampere University of Technology, Institute of Signal Processing, P.O.Box-553,
FIN-33101 Tampere, Finland
{karen, jta}@cs.tut.fi

³ IETR-TSI2M, UMR 6164 CNRS, University of Rennes I, 6, Rue de Kerampont,
B.P. 447, F-22305 Lannion Cedex, France
Benoit.Vozel@univ-rennes1.fr

Abstract. A standard sigma filter proposed by J.-S. Lee has found wide applications and frequent implementations in software packages. Later, several modifications have been introduced in order to improve its performance. In this paper we propose some new modifications trying to combine advantages of the original sigma and local statistic Lee filter as well as to ensure the filter robustness with respect to impulse noise. The basic performance characteristics of the proposed hybrid sigma filter are studied for cases of pure multiplicative noise. The comparison to some other well known filters is performed. A real life example of the designed filter application to side-look radar image is given.

Keywords: Image denoising, modified sigma filter, multiplicative noise.

1 Introduction

A standard sigma filter (SSF) [1] was proposed more than twenty years ago and since then it has found wide applications in processing of optical and radar images corrupted by Gaussian additive or multiplicative noise, respectively [2]. An evidence of the standard sigma filter popularity is that it has been implemented in many software packages intended on image processing needs (see, e.g., [3]). This is basically due to two main advantages of the SSF, namely, its excellent ability to preserve edges and fine details and its simplicity.

Based on advantages of the scalar SSF, its vector analogs have been designed [4], [5], [6]. Note that vector sigma filters possess the same advantages and they are very useful if a pre-processed multichannel or color image is a subject to further interpreting [4], [7]. Besides, the use of the SSF and its modifications has been shown expedient and effective within hard switching locally adaptive schemes [2], in particular, for solving a task of small sized object detection [8].

However, the SSF is characterized by several drawbacks restricting its applicability. The main among them is its rather poor noise suppression efficiency in image

homogeneous regions [9], [10]. Sometimes it is also desirable to provide filter ability to remove outliers and, respectively, to perform well in mixed noise environment.

Several modifications have been proposed in order to avoid aforementioned drawbacks of the SSF [9], [10], [11]. Due to them, the performance of the designed modified scalar sigma filters has become considerably better in image homogeneous regions and for mixed noise environment.

In this paper we propose one more modification of the SSF for which attractive features and operation principles of the SSF and the local statistic Lee filter [12] are successfully combined. The robustness to outliers can be also provided. The case of Gaussian multiplicative noise is considered. Extensive simulation data that allow comparing the characteristics for a set of filters are given. They confirm the efficiency of the proposed filter. Finally, an example of processing a real life image formed by side-look aperture radar (SLAR) is presented.

2 Hybrid Sigma Filter

Recall that in this paper we consider images for which multiplicative noise is the basic factor degrading their quality. Multiplicative noise is supposed to obey Gaussian distribution with mean equal to unity and (relative) variance $\sigma_\mu^2 < 0.1$. This is typical for remote sensing data obtained by SLARs or multilook synthetic aperture radars [2].

All sigma filters are based on one or another way of forming an initial neighbourhood for further determining scanning window pixel values that belong to this neighbourhood, and averaging of these values for output calculation. For the SSF, the neighbourhood forming is based on an assumption that the neighbourhood centre coincides with the noisy value of the scanning window central pixel $g(i, j)$:

$$y(i, j) = 1/N_\Delta(g(i, j)) \sum_{k,l=-M}^M \delta(k, l, g(i, j)) g(i+k, j+l), \quad (1)$$

where $N_\Delta(x) = \sum_{k,l=-M}^M \delta(k, l, x)$, $\delta(k, l, x) = \begin{cases} 1, & |g(i+k, j+l) - x| \leq x\Delta \\ 0, & |g(i+k, j+l) - x| > x\Delta \end{cases}$, Δ denotes

the averaging interval and it is commonly set equal to $2\sigma_\mu$ [1], $2(M+1)$ defines the scanning window size.

A rather low efficiency of noise suppression in image homogeneous regions observed for the SSF is explained by "incorrect" selection of neighbourhood centre in such cases when a noise value corrupting the scanning window central pixel relates to distribution "tail". Due to this, up to 60% of scanning window pixel values can be excluded from averaging, then it becomes inefficient.

The noise suppression efficiency in image homogeneous regions can be improved by modifying the way to select the neighbourhood centre g_s (for the SSF $g_s = g(i, j)$). Our proposition is the following. For the initial interval $[g(i, j) - \Delta g(i, j); g(i, j) + \Delta g(i, j)]$ let us find such g_h that the number of the given scanning window pixel values that belong to the interval $[g_h - \Delta g_h; g_h + \Delta g_h]$ is maximal

$$\forall g \in [g(i, j) - \Delta g(i, j); g(i, j) + \Delta g(i, j)] \quad N(g_h) \geq N(g) \quad (2)$$

To our experience [9], [10], such modified selection of a new neighborhood should lead to considerable improving the noise suppression efficiency in image homogeneous regions. However, this can also result in larger blurring. To partly alleviate this shortcoming, we propose one more modification. Let us take into account the local variance estimate in a scanning window similarly to that it is done for the local statistic Lee filter [12]. In opposite to the Lee's filter [12], let us calculate local variance in a scanning window with accounting not all values. This can be done as follows:

$$\sigma_h^2 = \left[1/N_h \sum_{k,l=-M}^M \delta_h(k,l) g(k,l)^2 - \bar{g}^2 \right] / \bar{g}^2, \quad \bar{g} = 1/N_h \sum_{k,l=-M}^M \delta_h(k,l) g(k,l), \quad (3)$$

$$\text{where } N_h = \sum_{k,l=-M}^M \delta_h(k,l), \quad \delta_h(k,l) = \begin{cases} 1, & g_{\min} \leq g(i+k, j+l) \leq g_{\max} \\ 0, & \text{otherwise} \end{cases},$$

$$g_{\min} = \min(g_h - \Delta g_h, g(i, j) - \Delta g(i, j)), \quad g_{\max} = \max(g_h + \Delta g_h, g(i, j) + \Delta g(i, j)).$$

With taking into consideration (1), (2), and (3), the output of the proposed hybrid sigma filter (HSF) is defined as

$$y_h(i, j) = \begin{cases} y_{imp}(i, j), & \sigma_h^2 \leq \sigma_\mu^2 \\ y_{imp}(i, j) + (1 - \sigma_\mu^2 / \sigma_h^2)(g(i, j) - y_{imp}(i, j)), & \sigma_h^2 > \sigma_\mu^2 \end{cases} \quad (4)$$

$$\text{where } y_{imp}(i, j) = 1/N_\Delta(g_h) \sum_{k,l=-M}^M \delta(k,l, g_h) g(i+k, j+l),$$

According to (4), the proposed HSF should perform better noise suppression if σ_h^2 is smaller, i.e., in image homogeneous regions or in edge/detail neighborhoods where the remained values (that belong to the new neighborhood) obey well to Gaussian distribution. On the contrary, if σ_h^2 is rather large, then $g(i, j)$ is taken into account with a larger weight to provide good preservation of edges and fine details.

If an image is corrupted by mixed noise, then we propose to apply the same procedure as for the modified sigma filters [9], [10]. It presumes the following: if for a given scanning window N_Δ for the original neighborhood is not larger than some present threshold N_{thres} , then one has to apply a robust detail preserving filter for obtaining HSF output. Otherwise, all operations (1) – (3) should be performed. As the robust detail preserving filter we recommend to use either 3LH+ variant of FIR hybrid median filter [13] if probability of impulse noise is less than 5% or center weighted median filter [14] with appropriately set parameters [15].

3 Analysis of Filter Performance for Artificial Test Images

For getting quantitative estimations of filters' performance, we have first used an artificially created test image presented in Fig. 1. This image contains a homogeneous region, edges, small sized and prolonged objects with positive and negative contrasts C (ratio of object intensities with respect to background, C is from 3 to 4).

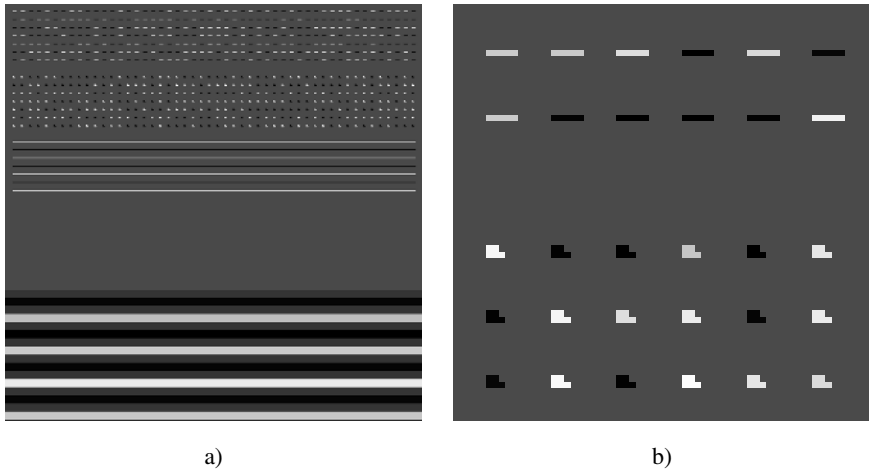


Fig. 1. The artificial test image (a) and its enlarged fragment that contains small sized objects

Consider the filter performance for the values of σ_μ^2 typical for radar images, namely, 0.006, 0.01, 0.02, 0.04, and 0.09. Table 1 contains the values of a parameter that characterizes noise suppression in image homogeneous regions. This parameter is

$$\theta = \sigma_{rem}^2 / \sigma_\mu^2, \quad (5)$$

where σ_{rem}^2 - is the residual noise variance in processed image

$$\sigma_{rem}^2 = 1/|G| \sum_{i,j \in G} [(g_{et}(i,j) - g(i,j)) / g_{et}(i,j)]^2, \quad (6)$$

where G is the considered area, $|G|$ denotes the number of pixels in this area, $g_{et}(i,j)$ denotes the ij -pixel value in the corresponding noise-free image.

The considered filters in Table 1 and other Tables below have the following notations: Sig5 and Sig7 are the SSFs with the scanning window sizes 5x5 and 7x7, respectively [1], Lee is the local statistic Lee filter (5x5) [12], MSig is the modified sigma filter (5x5) with $N_{thres} = 0$ (the parameter that denotes the number of pixels in the neighborhood for which the switching to FIR hybrid median filter is performed) [9], ASig denotes the Alparone's hybrid sigma filter [11], HSig5 and HSig7 are the proposed hybrid sigma filters with the windows 5x5 and 7x7, respectively.

The data presented in Table 1 confirm rather poor efficiency of noise suppression in image homogeneous regions for the SSF. Obviously, all modifications of the sigma filter as well as the Lee filter perform in such situations considerably (by 3...8 times) better. And the proposed filters HSig are among the best.

The integral values of θ determined in the edge neighbourhoods of the test image are given in Table 2. As seen in the sense of edge preservation the proposed HSF is also one of the best. It produces very good edge preservation if σ_μ^2 is small (or, in other words, if

Table 1. Efficiency of noise suppression in image homogeneous regions, θ

σ_{μ}^2	Sig5	Sig7	Lee	MSig	ASig	HSig5	HSig7
$\sigma_{\mu}^2 = 0.006$	0.2146	0.1808	0.0670	0.0608	0.1013	0.0578	0.0313
$\sigma_{\mu}^2 = 0.01$	0.2207	0.1881	0.0681	0.0642	0.0980	0.0599	0.0322
$\sigma_{\mu}^2 = 0.02$	0.2360	0.2033	0.0647	0.0707	0.1025	0.0582	0.0319
$\sigma_{\mu}^2 = 0.04$	0.2676	0.2347	0.0655	0.0941	0.1105	0.0624	0.0346
$\sigma_{\mu}^2 = 0.09$	0.3256	0.2964	0.0620	0.1457	0.1253	0.0789	0.0501

Table 2. Efficiency of noise suppression in edge neighbourhoods, θ

σ_{μ}^2	Sig5	Sig7	Lee	MSig	ASig	HSig5	HSig7
$\sigma_{\mu}^2 = 0.006$	0.2753	0.2257	0.9649	0.0831	0.1809	0.0917	0.0495
$\sigma_{\mu}^2 = 0.01$	0.2622	0.2140	0.9216	0.0933	0.1720	0.0897	0.0475
$\sigma_{\mu}^2 = 0.02$	0.2699	0.2256	0.9064	0.1086	0.1722	0.0924	0.0529
$\sigma_{\mu}^2 = 0.04$	0.3350	0.3049	0.8431	0.2045	0.2111	0.1363	0.1089
$\sigma_{\mu}^2 = 0.09$	0.5101	0.5059	0.7351	0.4529	0.3461	0.4977	0.5585

edges have enough high contrasts). Only for $\sigma_{\mu}^2=0.09$ the results for ASig become slightly better than for other analyzed filters.

Finally, Table 3 presents the integral values θ for the neighbourhoods of small sized objects in the artificial test image. As seen they can be very large for ASig filter if σ_{μ}^2 is small. This means that ASig considerably distorts small sized objects. Hsig produces the best results (the smallest θ) if σ_{μ}^2 is small. And only if $\sigma_{\mu}^2=0.09$ the proposed filter produces larger distortions in comparison to the SSF (compare the results for Sig5 and HSig5, for Sig7 and HSig7).

The aggregate preliminary conclusion is that the proposed HSF for the considered artificial test image and the analyzed set of σ_{μ}^2 produces an appropriate trade-off of the basic characteristics: efficiency of noise suppression in image homogeneous regions and edge/detail preservation.

Table 3. Efficiency of noise suppression in small sized object neighbourhoods, θ

σ_{μ}^2	Sig5	Sig7	Lee	MSig	ASig	HSig5	HSig7
$\sigma_{\mu}^2 = 0.006$	0.5650	0.5129	3.1858	0.4244	92.501	0.3488	0.2990
$\sigma_{\mu}^2 = 0.01$	0.5436	0.4826	4.2307	0.4035	52.569	0.3251	0.2695
$\sigma_{\mu}^2 = 0.02$	0.5515	0.5064	5.9393	0.4876	23.842	0.4818	0.4861
$\sigma_{\mu}^2 = 0.04$	0.7750	0.8138	8.5451	1.0288	10.225	0.7946	0.9224
$\sigma_{\mu}^2 = 0.09$	1.1481	1.3261	10.589	2.5447	3.4750	6.1620	9.7632

4 Analysis of Filter Performance for a Set of Standard Test Images and for Real Life SLAR Image

In the previous Section we have examined the filter performance for particular types of fragments of the artificial test image. Below, in Table 4, we give PSNR values for a set of standard test images: Baboon, Barbara, Goldhill, Lenna, and Peppers.

Table 4. Efficiency of noise removal for standard test images, PSNR, dB

Image	σ_μ^2	Sig5	Sig7	Lee	MSig	ASig	HSig5	HSig7
Baboon	$\sigma_\mu^2 = 0.006$	29.33	29.42	29.78	28.50	27.92	29.19	29.15
	$\sigma_\mu^2 = 0.01$	27.60	27.69	28.15	27.06	26.81	27.70	27.64
	$\sigma_\mu^2 = 0.02$	25.19	25.27	26.05	25.04	25.12	25.75	25.66
	$\sigma_\mu^2 = 0.04$	22.68	22.76	24.09	23.03	23.29	23.90	23.86
	$\sigma_\mu^2 = 0.09$	19.34	19.44	22.12	20.41	20.97	21.74	21.85
Barbara	$\sigma_\mu^2 = 0.006$	31.24	31.41	31.95	30.97	30.58	31.52	31.61
	$\sigma_\mu^2 = 0.01$	29.40	29.59	30.25	29.42	29.27	29.92	30.04
	$\sigma_\mu^2 = 0.02$	26.86	27.07	28.10	27.21	27.26	27.84	28.04
	$\sigma_\mu^2 = 0.04$	24.15	24.39	26.08	24.97	25.20	25.93	26.25
	$\sigma_\mu^2 = 0.09$	20.59	20.82	23.94	22.02	22.62	23.51	24.07
Goldhill	$\sigma_\mu^2 = 0.006$	31.88	31.89	32.70	32.02	32.19	32.63	32.41
	$\sigma_\mu^2 = 0.01$	30.20	30.23	31.32	30.81	30.90	31.38	31.13
	$\sigma_\mu^2 = 0.02$	27.69	27.75	29.49	29.00	28.99	29.69	29.49
	$\sigma_\mu^2 = 0.04$	24.95	25.07	27.83	27.00	26.95	28.01	27.98
	$\sigma_\mu^2 = 0.09$	21.18	21.35	25.79	23.63	24.11	25.32	25.67
Lenna	$\sigma_\mu^2 = 0.006$	32.25	32.35	33.79	33.32	33.28	33.90	33.73
	$\sigma_\mu^2 = 0.01$	30.37	30.51	32.31	31.97	31.80	32.55	32.48
	$\sigma_\mu^2 = 0.02$	27.69	27.90	30.42	30.09	29.66	30.79	30.87
	$\sigma_\mu^2 = 0.04$	24.62	24.87	28.44	27.51	27.19	28.72	29.08
	$\sigma_\mu^2 = 0.09$	20.60	20.82	26.08	23.40	23.90	25.49	26.18
Peppers	$\sigma_\mu^2 = 0.006$	32.35	32.45	33.69	33.47	33.34	33.92	33.80
	$\sigma_\mu^2 = 0.01$	30.58	30.75	32.41	32.29	32.03	32.77	32.74
	$\sigma_\mu^2 = 0.02$	27.86	28.10	30.57	30.43	29.96	31.07	31.24
	$\sigma_\mu^2 = 0.04$	24.81	25.09	28.68	27.77	27.56	28.98	29.45
	$\sigma_\mu^2 = 0.09$	20.80	21.05	26.42	23.69	24.28	25.74	26.52

The analysis of these data shows that PSNR values for the proposed HSF are practically always either the best or among the best. They are almost always by about 1 dB better than for MSig and HSig and at the same level as for the Lee filter [12].

The studies carried out in our papers [2] and [16] have demonstrated that Msig performs in the best manner for images that contain many relatively contrast edges and fine details like that one presented in Fig. 2,a.

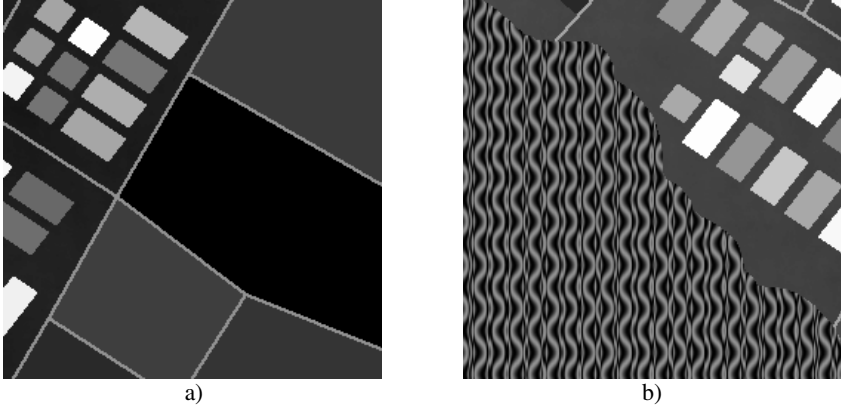


Fig. 2. Noise-free test images

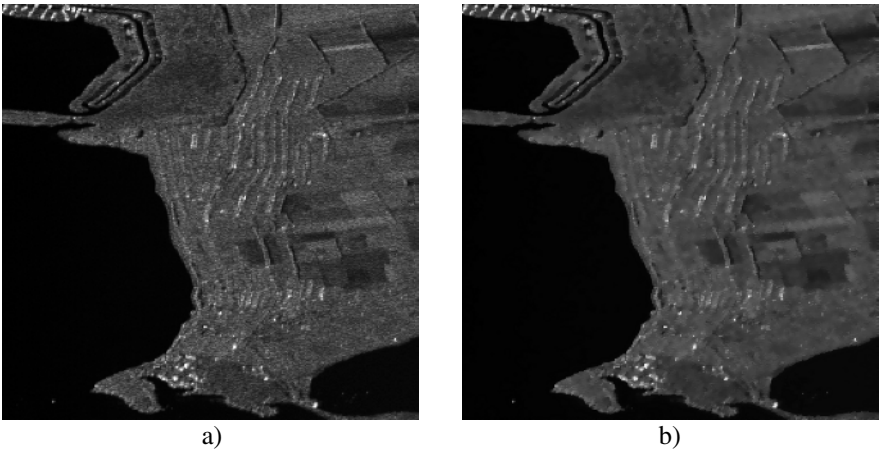


Fig. 3. Original SLAR X-band image (a) and the designed HSF output (b)

The same relates to the designed HSF. For such and similar images, the modified and designed hybrid sigma filters outperform other filters considered above by up to 4...6 dB if σ_{μ}^2 values are within the limits from 0.005 to 0.02. But for images containing a lot of textural regions like that one represented in Fig. 2, b as well as the test image Baboon MSF and HSF perform worse than some other filters. As shown in our paper [2], the best performance for such images is demonstrated by filters based on orthogonal transforms (in particular, the discrete cosine transform based filter [2]) or by more complex locally adaptive filters.

An example of real life image processing is shown in Fig. 3. This image was formed by an X-band airborne SLAR and kindly offered to us by A. Kalmykov Center for Earth Radiophysical Sensing. The value of σ_μ^2 estimated for this image by means of blind procedure [17] is about 0.012, i.e. just within the interval of σ_μ^2 simulated by us. Multiplicative noise presence in the original image (Fig. 3,a) is well observed in image homogeneous fragments with large mean intensities. The output of the proposed filter is depicted in Fig. 3,b. As seen, noise is efficiently suppressed while all small sized objects and edges are preserved well.

5 Conclusions

The carried out analysis of the designed hybrid sigma filter shows that it is able to outperform the standard and modified sigma filters as well as some other well known smoothers. In the sense of noise suppression efficiency, HSF approaches to the mean filter with the same scanning window size and is better than the standard median filter. Concerning edge/detail preservation, the proposed filter produces very good results for $\sigma_\mu^2=0.006\dots 0.04$, especially if edges and fine details are sharp and have rather large contrasts. Our HSF commonly outperforms earlier proposed modifications of the SSF.

Such properties allow recommending the HSF for processing of SLAR images contaminated by middle intensity multiplicative noise. This filter can be especially useful for pre-processing of images for which the primary task of their further interpreting consists in small object detection and localization.

Partly supported by the European Union. Co-financed by the ERDF and the Regional Council of Brittany, through the European Interreg3b PIMHAI project.

References

1. Lee, J.S.: Digital Image Smoothing and the Sigma Filter. *Computer Vision, Graphics and Image processing*, Vol. 24 (1983) 255-269
2. Tsymbal, O.B., Lukin, V.V., Ponomarenko, N.N., Zelensky, A.A., Egiazarian, K.O., Astola, J.T.: Three-state Locally Adaptive Texture Preserving Filter for Radar and Optical Image Processing. *EURASIP Journal on Applied Signal Processing*, No 8 (2005) 1185-1204
3. <http://www.rsinc.com/envi/>
4. Kurekin, A.A., Lukin, V.V., Zelensky, A.A., Tsymbal, O.V., Kulemin, G.P., Engman, E.T.: Processing multichannel radar images by modified vector sigma filter for soil erosion degree determination. *Proceedings SPIE/EUROPTO Symposium on Aerospace Remote Sensing*, SPIE Vol. 3868 (1999) 412-423
5. Lukac, R., Smolka, B., Plataniotis, K., Venetsanopoulos, A.: Generalized adaptive vector sigma filters. *Proceedings of International Conference on Multimedia and Expo*, Vol. 1 (2003) 537-540
6. Lukac, R., Smolka, B., Plataniotis, K.N., Venetsanopoulos, A.N.: Vector Sigma Filters for Noise Detection and Removal in Color Images. *Journal of Visual Communication and Image Representation*, Vol. 17, No 1 (2006) 1-26

7. Zelensky, A.A., Kulemin, G.P., Kurekin, A.A., Lukin, V.V., Tsymbal, O.V.: Modified Vector Sigma Filter for the Processing of Multichannel Radar Images and Increasing Reliability of Its Interpretation. *Telecommunications and Radioengineering*, Begell House (NY), Vol. 58, No 1-2 (2002) 100-113
8. Lukin, V., Ponomarenko, N., Zelensky, A., Astola, J., Egiazarian, K.: Automatic Design of Locally Adaptive Filters for Pre-processing of Images, Subject to Further Interpretation. *Proceedings of 2006 IEEE Southwest Symposium on Image Analysis and Interpretation*, (2006) 41-45
9. Lukin, V.V., Ponomarenko, N.N., Kuosmanen, P.S., Astola, J.T.: Modified Sigma Filter for Processing Images Corrupted by Multiplicative and Impulsive Noise. *Proceedings of EUSIPCO'96*, Vol. III (1996) 1909-1912
10. Lukin, V.V., Ponomarenko, N.N., Zelensky, A.A., Kurekin, A.A., Astola, J.T., Koivisto, P.T.: Modified Sigma Filter with Improved Noise Suppression Efficiency and Spike Removal Ability. *Proceedings of the 6-th International Workshop on Intelligent Signal Processing and Communication Systems*, (1998) 849-853
11. Alparone, L., Baronti, S., Garzelli, A.: A hybrid sigma filter for unbiased and edge-preserving speckle reduction. *Proceedings of International Geoscience and Remote Sensing Symposium*, (1995) 1409-1411
12. Lee, J.-S.: Speckle analysis and smoothing of synthetic aperture radar images. *Computer Vision, Graphics, Image Processing*, Vol. 17 (1981) 24-32
13. Heinonen, P., Neuvo, Y.: FIR-median hybrid filters. *IEEE Transactions on Acoustics, Speech, and Signal Processing*, Vol. 35 (1987) 832-838
14. Astola, J., Kuosmanen, P.: *Fundamentals of nonlinear digital filtering*. Boca Raton (USA): CRC Press LLC (1997)
15. Lukin, V.V., Koivisto, P.T., Ponomarenko, N.N., Abramov, S.K., Astola, J.T.: Two-stage methods for mixed noise removal. *Proceedings of International Workshop on Nonlinear Signal and Image Processing*, (2005) 128-133
16. Abramov, S.K., Lukin, V.V., Ponomarenko, N.N., Egiazarian, K., Pogrebnyak, O.B.: Influence of multiplicative noise variance evaluation accuracy on MM-band SLAR image filtering efficiency. *Proceedings of the Fifth International Kharkov Symposium Physics and Engineering of Millimeter and Sub-Millimeter Waves*, Vol. 1 (2004) 250-252
17. Abramov S.K., Lukin V.V., Zelensky A.A., Astola J.T.: Blind evaluation of noise variance in images using myriad operation. *Proceedings of IS&T/SPIE International Conference on Image Processing: Algorithms and Systems*, Vol. 4667 (2002) 192-203

Automatic Restoration of Old Motion Picture Films Using Spatiotemporal Exemplar-Based Inpainting*

Ali Gangal¹ and Bekir Dizdaroglu²

¹ Department of Electrical and Electronics Engineering, Karadeniz Technical University, 61080, Trabzon, Turkey
ali.gangal@ktu.edu.tr

² Program of Computer Technology and Programming, Besikduzu Vocational School, Karadeniz Technical University, Trabzon, Turkey
bekir@ktu.edu.tr

Abstract. This paper presents a method for automatic removal of local defects such as blotches and impulse noise in old motion picture films. The method is fully automatic and includes the following steps: fuzzy prefiltering, motion-compensated blotch detection, and spatiotemporal inpainting. The fuzzy prefilter removes small defective areas such as impulse noise. Modified bidirectional motion estimation with a predictive diamond search is utilized to estimate the motion vectors. The blotches are detected by the rank-ordered-difference method. Detected missing regions are interpolated by a new exemplar-based inpainting approach that operates on three successive frames. The performance of the proposed method is demonstrated on an artificially corrupted image sequence and on a real motion picture film. The results of the experiments show that the proposed method efficiently removes flashing and still blotches and impulse noise from image sequences.

1 Introduction

Vertical scratches and flashing blotches are artifacts commonly encountered in degraded motion picture films. These defective areas are caused either by accumulation of dirt or by the film material being abraded. They appear as bright and dark blotches on the scene, and are referred to as “dirt and sparkle” in the motion picture industry. Another type of defect is “still blotches”. Still blotches may be artificially formed to remove logos or undesired objects.

The successful restoration of corrupted image sequences involves mainly three processes: motion compensation, detection of blotches, and interpolation of the missing data in the blotched regions. In order to detect and interpolate the blotched regions, robust motion estimation techniques with respect to additive and replacement noise must be applied [1–3]. A priori information on the location of the missing regions facilitates the development of efficient algorithms to interpolate the missing

* This work was supported in part by the Research Foundation of Karadeniz Technical University under Grant 2004.112.004.01.

regions. Defects such as dirt and sparkle are usually determined by blotch detection methods, which use temporal discontinuities in the image sequence. Some existing methods for the detection of missing regions in image sequences are the spike detection index (SDI) method [4] and the rank-ordered-difference (ROD) detector [5,6]. Interpolation of the missing regions can be accomplished by methods based on a vector median filter (VMF), as presented in [7]. The VMF method promises good results, especially with respect to preserving detail and color. A common problem with detection and interpolation methods is that they yield poor performance when intense blotches and/or motion occur in successive image frames. It is difficult to interpolate useful data in areas of significant motion, such as motion involving rapid occlusion and uncovering. In this case, using more frames in the motion estimation and interpolation procedure may be a solution [7]. Recently, some techniques based on inpainting for removing objects from digital images or restoring damaged images have been presented in [8, 9]. It has been shown that large objects or damaged areas in an image can be successfully removed and considered, if the locations of the objects or damaged regions are known.

In this paper, we have extended the inpainting process to an interframe method, where the locations of the damaged regions are detected automatically. Our method is a hybrid restoration method that can remove flashing blotches, still blotches, and impulse noise. The proposed method not only combines the advantage of fuzzy filtering, motion-compensated blotch detection, and inpainting methods, but also improves on these methods when they are used to restore old motion picture films.

2 Proposed Method

Our proposed restoration method consists of a fuzzy prefilter, bidirectional motion estimation, automatic blotch detection with motion compensation, and spatiotemporal exemplar-based inpainting. Each stage used is represented in Fig. 1, and is described in the following subsections.

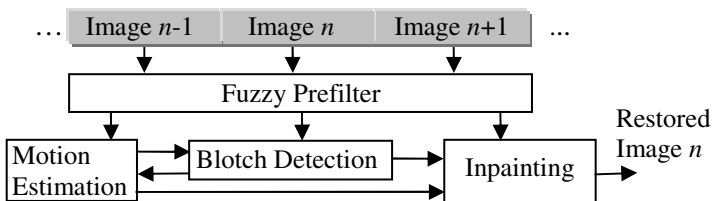


Fig. 1. Block diagram of the proposed restoration method

2.1 Fuzzy Prefilter

“Salt and pepper” noise may reduce the accuracy of the processes of motion estimation and inpainting. This type of noise can be removed successfully by intra-frame nonlinear filtering, applied to each frame separately. There are many filters

that perform well on various noisy images [10]; we chose a Gaussian fuzzy filter because of its good performance. We have modified this filter and applied it to color images. For a point (x, y) in an image, the filtered output is defined as follows:

$$\hat{I}^c(x, y) = \frac{\sum_{(i,j) \in A} I^c(x+i, y+j) \exp\left(-\left[I^c(x+i, y+j) - I_{med}^c\right]^2 / 2\sigma^2\right)}{\sum_{(i,j) \in A} \exp\left(-\left[I^c(x+i, y+j) - I_{med}^c\right]^2 / 2\sigma^2\right)}, \quad (1)$$

where $c = R, G, B$ represents the color components, A is the area of the window, I_{med}^c is the median value, and σ^2 is the value of the variance of $I^c(x, y)$ for $(x, y) \in A$.

2.2 Motion Estimation

Motion estimation plays an important role in most applications of image sequence processing, such as video compression, object tracking, and video restoration. In video restoration applications, some robust bidirectional motion estimation techniques have been developed and are already being used [1–3, 7]. These techniques are based mainly on block matching because of the ease of implementation. In these techniques, two or more sets of motion vectors are estimated dependently or independently. Then, motion trajectories that pass through the locations of missing pixels are estimated. These techniques give reasonable results for occlusion, uncovering, and scene cut problems. However, in corrupted image sequences, it is difficult to estimate motion trajectories accurately because blocks in large missing regions cannot be easily matched with blocks in the previous or subsequent frame. Therefore, the motion estimation process is adversely affected by the presence of large blotches.

We have used a bidirectional motion estimation method that is slightly different from the method presented in [7]. Since fuzzy filtering has been used, the multiresolution motion estimation used in [7] becomes unnecessary. For motion estimation, we used backward, forward, and cross searching applied to a single block, as shown in Fig. 2. This procedure can be implemented very easily and tends to be very robust in noisy and blotched regions. It consists of the following steps: (1) selecting a temporal mask involving three consecutive frames, in which the middle frame is the current frame to be restored; (2) dividing the frames into blocks; (3) searching the blocks in the previous and subsequent frames; (4) finding at most three candidates for the motion vector field; (5) postprocessing of the motion vectors to predict the final motion vector field for the current frame; and (6) shifting the mask by one frame and applying the above steps again. These steps are described below.

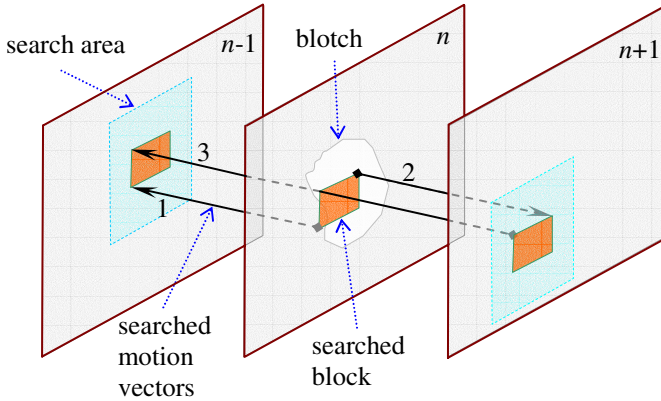


Fig. 2. Procedure for searching for motion vector

First, the frames are segmented into small rectangular blocks. Then, the blocks are searched according to the procedure shown in Fig. 2. In this figure, the beginning and the end of each arrow show the reference frame and the searched frame, respectively. Each block of a frame requires three independent searches to find the motion vectors. The first search is from the n th frame to the $(n - 1)$ th frame, the second search is from the n th frame to the $(n + 1)$ th frame, and the third search is from the $(n + 1)$ th frame to the $(n - 1)$ th frame. To achieve fast block matching, we used a predictive diamond search algorithm [11]. In this algorithm, the center of the search area is the predicted motion vector (the motion vector of the previous block). A large diamond pattern consisting of nine points is then examined. If the minimum is found to be the same as the prediction, then a smaller diamond is used and the final four points around the prediction are examined to optimize the final motion vector. If the minimum is not found to be the same as the prediction, the search moves back to the initial center, and the original diamond search algorithm is used until the best match is found in the center of the diamond pattern.

The searching approach illustrated by arrows 1 and 2 in Fig. 2 is bidirectional and therefore this approach can solve problems of covered and uncovered areas, and of scene cuts. This approach also improves the success of the matching.

If there is a large blotch, as big as an entire block in the current frame, however, the approach described above will not provide good results. This problem is significantly reduced by using an approach corresponding to arrow 3 in Fig. 2. With this approach, a motion vector trajectory passing through the block in the blotched area can be estimated more accurately. On the assumption of constant velocity, the motion vectors estimated by the searching approach illustrated by arrow 3 will pass through the location $\vec{r} + (\vec{v}_{n-1, n+1}(\vec{r}))/2$ in the current frame (here, $\vec{r} = (x, y)$, and $\vec{v}(\cdot)$ is the motion vector). This location may not coincide with the actual center of the block.

After checking whether occluded motion vectors have been obtained by these searches, the motion vector which has the minimum matching error is selected. This procedure is repeated for each pixel in the block. Pixels with no motion vectors are represented by a single vector that is obtained by averaging the motion

vectors associated with the other pixels. This procedure is applied if at least half of the pixels have a motion vector and if the maximum angle between the predicted vectors is less than a predefined value. The estimated motion vectors, along with the related frame numbers, are provided to the blotch detection and inpainting stages described in Section 2.3 and 2.4, respectively.

2.3 Blotch Detection

The blotch detector utilized in our method is based on the ROD method because that method is simple and has high detection performance. In the following, we describe our method.

Let $I_n(\vec{r})$ be the intensity value of the pixel at the spatial coordinates $\vec{r} = (x, y)$ in the n th frame. Let \mathbf{P} be a vector defined by

$$\mathbf{P} = [p_1, p_2, \dots, p_6], \quad (2)$$

where the pixel values p_1, p_2, \dots, p_6 are taken from the motion-compensated previous and subsequent frames at locations that are spatial neighbors of $I_n(\vec{r})$, as shown in Fig. 3.

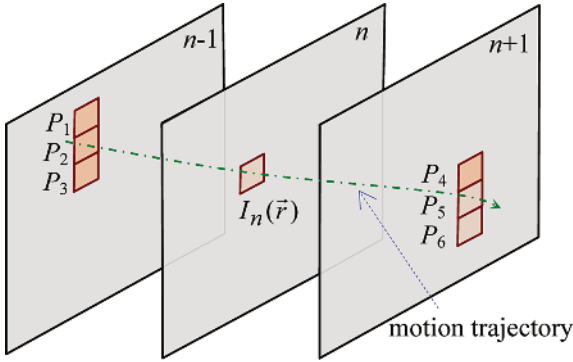


Fig. 3. Blotch detection procedure

Blotch detection is done according to the following steps. First, the elements of \mathbf{P} are ordered by rank, which yields a vector \mathbf{R} ,

$$\mathbf{R} = [R_1, R_2, \dots, R_6], \quad (3)$$

where $R_1 \leq R_2 \leq \dots \leq R_6$. The rank-order mean is $m = (R_3 + R_4) / 2$. Three rank-order differences are then calculated for \mathbf{R} . These differences are denoted by $D_l(\vec{r})$, $l = 1, 2, 3$, and defined as follows:

$$D_l(\vec{r}) = \begin{cases} R_l - I_n(\vec{r}) & \text{if } I_n(\vec{r}) \leq m \\ I_n(\vec{r}) - R_{7-l} & \text{if } I_n(\vec{r}) > m \end{cases} \quad (4)$$

The pixel at the location $\vec{r} = (x, y)$ in the frame n is detected as corrupted if at least one of the rank-order differences exceeds one of the preselected thresholds T_l , which determine the detector's sensitivity. The threshold values were selected such that $0 < T_1 < T_2 < T_3$.

2.4 Spatiotemporal Exemplar-Based Inpainting

We propose a spatiotemporal exemplar-based inpainting method for the restoration of damaged regions of the image. This method is like the image inpainting method presented in [8] for still images. In [8], Criminisi et al. proposed an exemplar-based inpainting method, which fills in the target region with patches from the source region possessing a similar texture. The candidate patches are selected from the whole image, with special priority being given to those along the isophotes (lines of equal gray value) so as to preserve the linear structure during the filling-in. Criminisi's method is intraframe and considers blotched areas to be known. So, if the missing regions cannot be found exactly from undamaged areas of the current frame, the restoration may fail. The major disadvantage of this method is the global searching, which not only leads to errors in the match but also greatly decreases system performance. It merely adopts a simple priority computing strategy without considering the cumulative matching error. Therefore we have extended the method into three dimensions. Our method uses the points where the motion trajectories pass through the frames as starting points for the search algorithm. Therefore it searches accurately and rapidly in a small region. The algorithm uses three sequential frames at the same time. The proposed spatiotemporal inpainting process is shown in Fig. 4. The steps of the algorithm are as follows:

Step 1. Determine the blotched area to be used as the target region. Compute the number of pixels of the target region to be used for setting up the size of the patches on the contour $\delta\Omega$ of the target region Ω . Repeat steps 2–7 below for every blotched area while $\Omega > 0$.

Step 2. Identify the fill front $\delta\Omega^t$. If $\Omega^t = 0$, exit (the superscript t indicates the current iteration).

Step 3. Determine the source region for each blotched area in the input image, as shown in Fig. 4.

Step 4. Determine the filling priorities so that the method to be capable of propagating information about both texture and structure from the exemplar-based filling algorithm. It performs the synthesis task through a best-first filling strategy that depends entirely on the priority values that are assigned to each patch on the fill front. The computation of the priorities is biased toward those patches which are on the continuations of strong edges and are surrounded by high-confidence pixels. Given a patch $\psi_{\vec{r}}$ centered at a point \vec{r} on the contour $\delta\Omega$ of the target region Ω , its priority $P(\vec{r})$ (i.e. $P(\vec{r}) = \forall \vec{r} \in \delta\Omega^t$) is defined as

$$P(\vec{r}) = C(\vec{r})D(\vec{r}), \quad (5)$$

where $C(\vec{r})$ and $D(\vec{r})$ are called the confidence term and the data term, respectively. The confidence term is a measure of the amount of reliable information surrounding the pixel \vec{r} . The intention is to fill first those patches which have more of their pixels already filled, with additional preference given to pixels that were filled early on (or were never part of the target region). The priority $P(\vec{r})$ is computed for every patch centered on the contour $\delta\Omega$. The confidence term $C(\vec{r})$ is defined by

$$C(\vec{r}) = \frac{1}{A_{\psi_{\vec{r}}}} \sum_{\vec{q} \in \psi_{\vec{r}} \cap \Phi_n} C(\vec{q}), \quad (6)$$

where $A_{\psi_{\vec{r}}}$ is the area of $\psi_{\vec{r}}$, Φ_n is the source region in the n th frame, \vec{q} is the vector of a pixel point inside $\psi_{\vec{r}}$, and $C(\vec{q})$ is the previously calculated confidence term at the pixel point \vec{q} . For initialization, the confidence term $C(\vec{r})$ is set to 0 for the target region Ω and set to 1 for the source region Φ_n .

The data term $D(\vec{r})$ is a function of the strength of the isophotes hitting the front $\delta\Omega$. This factor encourages linear structures to be synthesized first, and therefore propagated securely into the target region. The data term $D(\vec{r})$ is defined as

$$D(\vec{r}) = |\nabla I_{\vec{r}}^{\perp} \cdot \mathbf{n}_{\vec{r}}| / 255, \quad (7)$$

where, $\mathbf{n}_{\vec{r}}$ is the unit vector orthogonal to the front $\delta\Omega$ at the point \vec{r} , and \perp denotes the orthogonal operator. The gradient $\nabla I_{\vec{r}}^{\perp}$ at the pixel point \vec{r} and the unit vector $\mathbf{n}_{\vec{r}}$ are defined as

$$\nabla I_{\vec{r}}^{\perp} = [(I(x, y+1) - I(x, y-1))/2 \quad -(I(x+1, y) - I(x-1, y))/2], \quad (8)$$

$$\mathbf{n}_{\vec{r}} = 1/\|\mathbf{n}_{\vec{r}}\| [n(x+1, y) - n(x, y) \quad n(x, y+1) - n(x, y)]^T, \quad (9)$$

where T indicates the transpose, and

$$\|\mathbf{n}_{\vec{r}}\| = \sqrt{[n(x+1, y) - n(x, y)]^2 + [n(x, y+1) - n(x, y)]^2}. \quad (10)$$

If the location of $n(\cdot)$ falls into the target region, its value set to 0; otherwise, it is set to 1.

Step 5. Once all priorities on the fill front have been computed, the patch $\psi_{\hat{\vec{r}}}$ with the highest priority is found, i.e., $\hat{\vec{r}} = \arg \max_{\vec{r} \in \delta\Omega} P(\vec{r})$. This patch is then filled with data extracted from the source regions Φ_{n-1} , Φ_n , and Φ_{n+1} . The centers of the

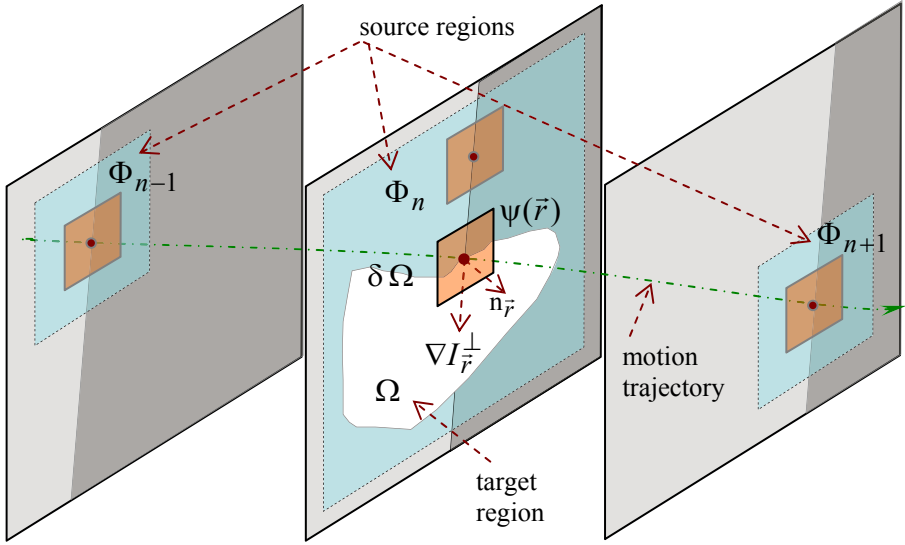


Fig. 4. Spatiotemporal exemplar-based inpainting process

search regions are taken to be on the motion trajectory for all of the frames. The image point which the motion trajectory passes through is taken as only the initial search point for the patches, because the center of a patch may not coincide with the actual center of the matched block in the motion estimation process.

Step 6. The image texture is propagated by direct sampling of the source region. The patch which is most similar to $\psi_{\hat{r}}$ is searched for in the source regions to obtain the source exemplar

$$\psi_{\hat{q}} = \arg \min_{\psi_{\hat{q}} \in [\Phi_{n-1}, \Phi_n, \Phi_{n+1}]} d(\psi_{\hat{r}}, \psi_{\hat{q}}), \quad (11)$$

where the distance $d(\cdot)$ between two generic patches is defined simply as the sum of the squared differences for the pixels already filled in the two patches.

Step 7. Once the source exemplar $\psi_{\hat{q}}$ has been found, the image data is copied from $\psi_{\hat{q}}$ to $\psi_{\hat{r}} \forall \vec{r} \in \psi_{\hat{r}} \cap \Omega$. After the patch $\psi_{\hat{r}}$ has been filled with new pixel values, $C(\vec{r})$ is updated in the area delimited by $\psi_{\hat{r}}$: $C(\vec{r}) = C(\hat{\vec{r}}), \forall \vec{r} \in \psi_{\hat{r}} \cap \Omega$.

Since the centers of the source regions in the previous and subsequent frames are taken to be on the motion trajectories, the search areas within these frames can be smaller than those in the current frame. Hence this inpainting method spends less time on searching in the previous and subsequent frames than it does in the n th frame.

3 Experimental Results

The validity of the proposed method was tested on both an artificially corrupted and a real image sequence that contained flashing and still blotches. The quantitative and qualitative performance of the method was compared with that of the methods presented in [7, 8]. The size of the images in both sequences was 352×288 pixels. The block size was chosen as 4×4 for the block matching. The patch size was taken as maximum 9×9 and minimum 5×5 for the exemplar-based inpainting. The size of the source regions was 50×50 in the current frame and 16×16 in the other frames.

We used 400 frames of an image sequence called the ‘‘Foreman’’ sequence. The sequence was corrupted by artificially generated flashing blotches (5%), still blotches (2%), and impulse noise (10%). Some of the blotches were placed in regions that included rapid motion, occlusion, and uncovering. Three typical corrupted consecutive images (with frame numbers 254, 255, and 256) from the ‘‘Foreman’’ sequence are shown in Fig. 5(a). As can be seen from this figure, a still blotch has been added to the upper right corner of the sequence, which might represent a television logo. Another sequence used was a real image sequence from a motion picture film, captured from a television broadcast. Frames 51, 52 and 53 of this sequence are shown in Fig. 5(b).

3.1 Performance Analysis on the Artificially Corrupted Image Sequence

To evaluate the restoration performance of the method, we computed the normalized mean squared error (NMSE); this is a standard quantitative measure, defined by

$$\text{NMSE} = \frac{\sum_{(x,y) \in \mathbf{I}} \sum_{c=R,G,B} \left(I^c(x,y) - \tilde{I}^c(x,y) \right)^2}{\sum_{(x,y) \in \mathbf{I}} \sum_{c=R,G,B} \left(I^c(x,y) \right)^2}, \quad (12)$$

where $\tilde{I}^c(\cdot)$ is the restored pixel value, and \mathbf{I} is the entire image. A graph of the NMSE versus frame number for frames 250–270 (21 frames) of the ‘‘Foreman’’ sequence is shown in Fig. 6. As can be seen from this figure, the proposed method shows a smaller NMSE and less error variation than the do other two methods, even for heavily corrupted frames. We used a fuzzy prefilter with the methods described in [7, 8] also, in order to compare the methods under the same conditions.

In a second experiment, we evaluated the methods qualitatively. We chose the values of T_1 , T_2 , and T_3 as 37, 39, and 55, respectively, for the ROD detector. The results for the 255th frame of the ‘‘Foreman’’ sequence are shown in Fig. 7. As can be seen from Fig. 7(b), the still blotches were removed successfully by the method presented in [8] (the region shown by the dashed rectangle), but in the region that shows a hand moving (shown by the solid rectangle) the method has failed. There was no blotch detection here, because in this method the blotches are assumed to be known. As shown in Fig. 7(c), the region containing the moving hand



Fig. 5. Three consecutive frames taken from (a) artificially corrupted “Foreman” sequence (b) an old motion picture film

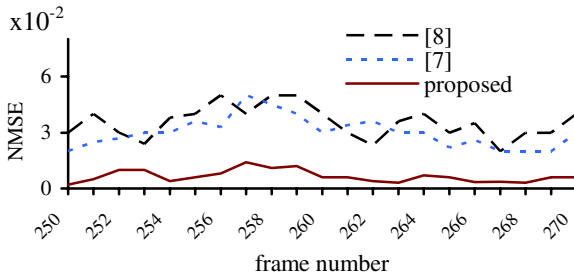


Fig. 6. NMSE for “Foreman” sequence artificially corrupted by flashing blotches (5%), still blotches (2%), and salt and pepper noise (10%)

(dotted rectangle) was restored successfully by the method presented in [7]. However, the still blotch (the region shown by the solid rectangle) could not be removed because it was not detected. The blotches were detected automatically here. Figure 7(d) shows that all of the critical regions, containing rapid movement and a still blotch, were successfully restored by the proposed method.

3.2 Performance Analysis on the Real Image Sequence

We applied the same methods to the degraded real image sequence shown in Fig. 5(b). The detected blotches are shown in Fig. 8(a) for the 52nd frame. The results of the methods of [8] and [7] and of the proposed method are shown in Figs. 8(b), (c), and (d), respectively. As can be seen from the figure, the blotched regions in the rectangles were successfully restored by the proposed method.

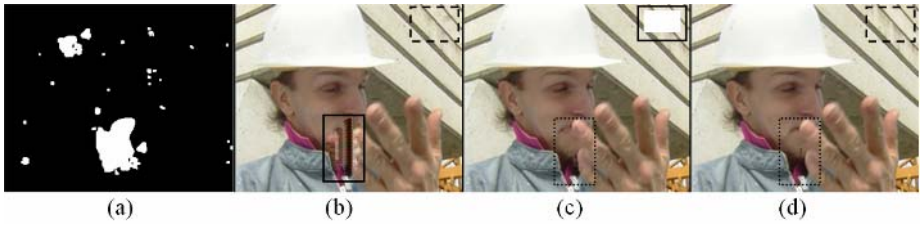


Fig. 7. A sample of the result of restoration of the artificially corrupted “Foreman” sequence. (a) Detected blotches, (b) the result of the method presented in [8], (c) the result of the method presented in [7], and (d) the result of the proposed method. The solid, dotted, and dashed rectangles indicate that the results in those critical areas are poor, moderate, and good, respectively.



Fig. 8. A sample of the result of restoration of a real corrupted image sequence (frame number 52). (a) Detected blotches, (b) the result of the method presented in [8], (c) the result of the method presented in [7], and (d) the result of the proposed method. The solid rectangles indicate regions in which the restoration has failed, and the dashed rectangles show the regions in which restoration was successful.

The methods were implemented in Visual C++.NET without regard to computational efficiency, and run on a Pentium 2.4 GHz computer with 256 Mbytes of physical memory. The average time required to restore a frame depends on the size of the blotted regions, the size of the images, and the amount of motion. For the corrupted image sequence shown in Fig. 5(b), the average process times of the proposed method, the method of [7], and the method of [8] were 116 seconds/frame, 72 seconds/frame, and 59 seconds/frame, respectively.

4 Conclusions

In this paper, we have proposed a new method for the restoration of motion picture films that have deteriorated owing to flashing blotches, still blotches, and impulse noise. The main elements of the method are fuzzy prefiltering, modified bidirectional motion estimation, blotch detection, and interpolation of missing data by spatiotemporal exemplar-based inpainting.

The performance of the method was tested not only on an artificially corrupted image sequence but also on a naturally degraded video of an old motion picture film, and as compared with that of the methods presented in [7, 8]. The results show not only that the method automatically detects and removes flashing blotches successfully, but also that it

successfully removes still blotches and impulse noise because it combines the advantages of the methods of [7, 8] for the interpolation of such regions.

References

1. Thoma, R., and Bierling, M.: Motion compensating interpolation considering covered and uncovered background. *Signal Processing: Image Commun.*, Vol. 1(2) (1989) 191–212
2. Kim, M.K., and Kim, J.K.: Efficient motion estimation algorithm for bidirectional prediction scheme. *IEE Electron. Lett.*, Vol. 30(8) (1994) 632–633
3. Goh, W.B., Chong, M.N., Kalra, S., and Krishnan, D.: Bi-directional 3D auto-regressive model approach to motion picture restoration. *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing*. Atlanta, USA, (1996) 2275–2278
4. Kokaram, A.C., Morris, R.D., Fitzgerald, W.J., and Rayner, J.W.: Detection of missing data in image sequences. *IEEE Trans. Image Processing*, Vol. 4(11) (1995) 1496–1508
5. Nadenau, M.J., and Mitra, S.K.: Blotch and scratch detection in image sequences based on rank ordered differences. *Proceedings of the 5th International Workshop on Time-Varying Image Processing and Moving Object Recognition*, Florence, Italy (1996) 1–7
6. Armstrong, S., Kokaram, A.C., and Rayner, P.J.W.: Restoring video images taken from scratched 2-inch tape. In *Workshop on Non-Linear Model Based Image Analysis (NMBIA'98)*, eds. S. Marshall, N. Harvey, D. Shah, Springer, Verlag (1998) 83–88
7. Gangal, A., Kayıkcıoglu, T., and Dizdaroglu, B.: An improved motion-compensated restoration method for damaged color motion picture films. *Signal Processing: Image Commun.*, Vol. 19 (2004) 353–368
8. Criminisi, A., Perez, P., and Toyama, K.: Region filling and object removal by exemplar-based inpainting. *IEEE Trans. Image Processing*, Vol. 13(9) (2004) 1200–1212
9. Bertalmio, M., Vese, L., and Sapiro, G.: Simultaneous structure and texture image inpainting. *IEEE Trans. Image Processing*, Vol. 12(8) (2003) 882–889
10. Kwan, H.K.: Fuzzy filters for noisy image filtering. *Proceedings of the 2003 International Symposium on Circuits and Systems (ISCAS'03)*, Vol. 4, (2003) 161–164.
11. Tourapis, A.M., Shen, G., Liou, M.L., Au, O.C., and Ahmad, I.: A new predictive diamond search algorithm for block based motion estimation. *Proceedings of Visual Communication and Image Processing*, eds. K. N. Ngan, T. Sikora, M. T. Sun, Perth, Australia, (2000) 1365–1373.

Dedicated Hardware for Real-Time Computation of Second-Order Statistical Features for High Resolution Images

Dimitris Bariamis, Dimitris K. Iakovidis, and Dimitris Maroulis

Dept. of Informatics and Telecommunications, University of Athens,
Panepistimiopolis, Illisia, 15784 Athens, Greece
rtsimage@di.uoa.gr

Abstract. We present a novel dedicated hardware system for the extraction of second-order statistical features from high-resolution images. The selected features are based on gray level co-occurrence matrix analysis and are angular second moment, correlation, inverse difference moment and entropy. The proposed system was evaluated using input images with resolutions that range from 512×512 to 2048×2048 pixels. Each image is divided into blocks of user-defined size and a feature vector is extracted for each block. The system is implemented on a Xilinx VirtexE-2000 FPGA and uses integer arithmetic, a sparse co-occurrence matrix representation and a fast logarithm approximation to improve efficiency. It allows the parallel calculation of sixteen co-occurrence matrices and four feature vectors on the same FPGA core. The experimental results illustrate the feasibility of real-time feature extraction for input images of dimensions up to 2048×2048 pixels, where a performance of 32 images per second is achieved.

1 Introduction

The second-order statistical information present in an image relates to the human perception of texture. It has been successfully utilized in a variety of machine vision systems, including biomedical [1,2], remote sensing [3], quality control [4], and industrial defect detection systems [5].

A well established statistical tool that captures the second-order statistical information is the co-occurrence matrix [6]. The calculation of the co-occurrence matrix has a complexity of only $O(N^2)$ for an input image of $N \times N$ -pixel dimensions, but the calculation of multiple matrices per time unit increases the processing power requirements. Using software co-occurrence matrix implementations running on conventional general-purpose processors does not enable real-time performance in a variety of applications, which require a high number of calculated matrices per time unit. Such demanding applications in the field of image processing include analysis of video streams [1,6], content-based image retrieval [7], real-time industrial applications [5] and high-resolution multispectral image analysis [2].

Field Programmable Gate Arrays (FPGAs) are high-density reconfigurable devices that can be hosted by conventional computer hardware [9]. They enable the rapid and

low cost development of circuits that are adapted to specific applications and exploit the advantages of parallel architectures. A dedicated hardware system that efficiently computes co-occurrence matrices in parallel can meet the requirements for real-time image analysis applications. The Very Large Scale Integration (VLSI) architectures [10] provide an alternative to the FPGAs, but have drawbacks such as higher cost and time-consuming development. Furthermore, they cannot be reconfigured.

Within the first FPGA-based systems dedicated to co-occurrence matrix computations, was the one presented in [5,11]. It involves the computation of two statistical measures of the co-occurrence matrix. Moreover, the measures are extracted indirectly, without calculating the co-occurrence matrix itself. A later work by Tahir et al. [2] presents an FPGA architecture for the parallel computation of 16 co-occurrence matrices. The implementation exploits the symmetry, but not the sparseness of the matrices, resulting in a large FPGA area utilization. This leads to the need of a separate core for the feature calculation. Thus, the system is capable of processing high-resolution images, but does not achieve real time performance.

In this paper, we present a novel FPGA based system that allows the parallel computation of 16 co-occurrence matrices and 4 feature vectors. The dedicated hardware exploits both the symmetry and the sparseness of the co-occurrence matrix and uses an efficient approximation method for the logarithm, enabling real-time feature extraction for input images of dimensions up to 2048×2048 pixels. Furthermore, the system comprises of a single core for both the co-occurrence matrix and the feature calculation. Thus, no overhead is incurred by reprogramming cores onto the FPGA in order to calculate the feature vectors.

The paper is organized in five sections. Section 2 refers to the second-order statistical features and their integer arithmetic formulation. The architecture of the proposed system is described in Section 3. Section 4 presents the experimental results that demonstrate the system performance. The conclusions of this study are summarized in Section 5.

2 Second-Order Statistical Features

The co-occurrence matrix of an $N \times N$ -pixel image block, encodes the gray-level spatial dependence based on the estimation of the second-order joint-conditional probability density function $P_{d,\theta}(i, j)$. It is computed by counting all pairs of pixels of an image block at distance d having gray-levels i and j at a given direction θ .

$$P_{d,\theta}(i, j) = \frac{C_{d,\theta}(i, j)}{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} C_{d,\theta}(i, j)} \quad (1)$$

where $C_{d,\theta}(i, j) = \# \{(m, n), (u, v) \in N \times N: f(m, n) = j, f(u, v) = i, |(m, n) - (u, v)| = d, \angle((m, n), (u, v)) = \theta\}$, $\#$ denotes the number of elements in the set, $f(m, n)$ and $f(u, v)$ correspond to the gray-levels of the pixel located at (m, n) and (u, v) respectively, and N_g is the total number of gray-levels in the image [6]. We choose $N_g = 32$ (5-bit representation).

The co-occurrence matrix can be regarded symmetric if the distribution between opposite directions is ignored. The symmetric co-occurrence matrix is derived as $P_{d,\theta}(i, j) = (P_{d,\theta}(i, j) + P_{d,\theta}(j, i))/2$. Therefore, the co-occurrence matrix can be represented as a triangular structure without any information loss, and θ is chosen within the range of 0° to 180° . Common choices of θ include 0° , 45° , 90° and 135° [1,2,6,12].

Moreover, depending on the image dimensions, the co-occurrence matrix can be very sparse, as the number of gray-level transitions for any given distance and direction, is bounded by the number of image pixels.

Out of the 14 features originally proposed by Haralick et al. [6] we have considered four, namely angular second moment (f_1), correlation (f_2), inverse difference moment (f_3) and entropy (f_4):

$$f_1 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P_{d,\theta}^2(i, j) \quad (2)$$

$$f_2 = \frac{\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} i \cdot j \cdot P_{d,\theta}(i, j) - \mu_x \cdot \mu_y}{\sigma_x \cdot \sigma_y} \quad (3)$$

$$f_3 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} \frac{1}{1+(i-j)^2} P_{d,\theta}(i, j) \quad (4)$$

$$f_4 = -\sum_{i=1}^{N_g} \sum_{j=1}^{N_g} P_{d,\theta}(i, j) \cdot \log P_{d,\theta}(i, j) \quad (5)$$

where μ_x , μ_y , σ_x and σ_y are the means and the standard deviations of the marginal probabilities $P_x(i)$ and $P_y(j)$ obtained by summing the rows and columns of matrix $P_{d,\theta}(i, j)$ respectively. These four measures have been shown to provide high discrimination accuracy that can only be marginally increased by adding more features to the feature vector [1], [13].

The calculation of the four measures requires floating point operations that result in higher FPGA area utilization and lower operating frequencies. To implement the calculation of the measures efficiently in hardware, we have extracted five expressions that can be calculated using integer arithmetic:

$$V_1 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} C_{d,\theta}^2(i, j) \quad (6)$$

$$V_2 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} i \cdot j \cdot C_{d,\theta}(i, j) \quad (7)$$

$$V_3 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} C_{d,\vartheta}(i, j) \cdot IDMLUT[|i-j|] \quad (8)$$

$$V_4 = \sum_{i=1}^{N_g} \sum_{j=1}^{N_g} C_{d,\vartheta}(i, j) \cdot (2^{28} \cdot \log C_{d,\vartheta}(i, j)) \quad (9)$$

$$V_5 = \sum_{i=1}^{N_g} C_x^2(i) \quad (10)$$

$$C_x(i) = r \cdot P_x(i) \quad (11)$$

$$r = \sum \sum C_{d,\vartheta}(i, j) \quad (12)$$

$$IDMLUT[k] = \left\lfloor \frac{2^{31}}{1+k^2} \right\rfloor \quad (13)$$

The logarithm in Eq. (9) is approximated using the method described in Section 3.2, whereas *IDMLUT* is a 32×32-bit Look Up Table (LUT) used for the calculation of the Inverse Difference Moment. The result of the calculation in hardware is a vector $\bar{V} = [V_1, V_2, V_3, V_4, V_5]$ that is used for the calculation of the four Haralick features through the use of the following equations:

$$f_1 = \frac{V_1}{r^2} \quad (14)$$

$$f_2 = \frac{(N_g - 1) \cdot (r \cdot N_g^2 \cdot V_2 - r^2)}{N_g^2 \cdot V_5 - N_g \cdot r^2} \quad (15)$$

$$f_3 = \frac{V_3}{2^{31} \cdot r} \quad (16)$$

$$f_4 = r \log r - \frac{V_4}{2^{28} \cdot r} \quad (17)$$

Eqs. (14)-(17) are executed in software. The computation of these equations incurs a negligible overhead to the overall system performance.

3 System Architecture

The architecture of the proposed system was developed in Very High Speed Integrated Circuits Hardware Description Language (VHDL). It was implemented on a Xilinx Virtex-XCV2000E-6 FPGA, which is characterized by 80×120 Configurable

Logic Blocks (CLBs) providing 19,200 slices (1 CLB = 2 slices). The device includes 160 256×16-bit Block RAMs and can support up to 600kbit of distributed RAM. The host board, Celoxica RC-1000 has four 2MB static RAM banks. The FPGA and the host computer can access the RAM banks independently, whereas onboard arbitration and isolation circuits prohibit simultaneous access.

The system architecture is illustrated in Fig. 1. The FPGA implementation includes a control unit, four memory controllers (one for each memory bank), 16 Co-occurrence Matrix Computation Units (CMCUs) and four Vector Calculation Units (VCUs). Each input image is divided into blocks of user-specified dimensions and loaded into a corresponding RAM bank using a 25-bit per pixel representation. Each pixel is represented by a vector $\vec{a} = [a_p, a_0, a_{45}, a_{90}, a_{135}]$ that comprises of five 5-bit components, namely, the gray-level a_p of the pixel and the gray-levels a_0, a_{45}, a_{90} and a_{135} of its neighboring pixels at $0^\circ, 45^\circ, 90^\circ$ and 135° directions.

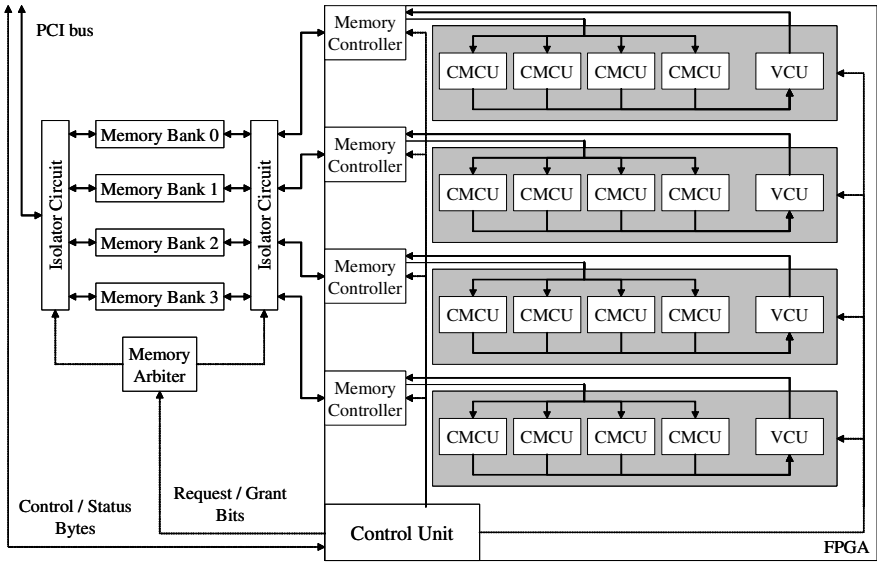


Fig. 1. Overview of the system architecture

All FPGA functions are coordinated by the control unit, which generates synchronization signals for the memory controllers, the CMCUs and the VCUs. The control unit also handles communication with the host, by exchanging control and status bytes, and requesting or releasing the ownership of the on-card memory banks. The system includes 16 CMCUs that are grouped in four quadruplets. Each CMCU in the quadruplet reads the vectors \vec{a} that represent an image block from one of the memory controllers and computes the GLCM for a single direction. The 16 CMCU outputs of the four quadruplets are connected to the four VCUs that calculate the vectors \vec{V} from the GLCMs. These vectors are written to the on-card memory through the memory controllers.

3.1 Co-occurrence Matrix Computation Units

Considering the requirements of the proposed application, the CMCU was developed to meet three main objectives: small FPGA area utilization, high throughput per clock cycle and high frequency potential. The small area utilization allows the implementation of the four VCUs on the same core, whereas the high throughput and frequency ensure the high efficiency of the design. To meet these three objectives we have considered various alternatives for the implementation of the CMCUs. These include the utilization of the existent FPGA BlockRAM arrays, the implementation of standard sparse array structures that store pairs of indices and values, and the implementation of set-associative [14] sparse arrays. The BlockRAM arrays and the standard sparse array structures would not suffice to meet all three objectives. The BlockRAM arrays would lead to larger area utilization, compared with the sparse implementations, whereas the standard sparse arrays would result in a lower throughput, compared with the other implementations, as the cycles needed to traverse the indices of the array are proportional to its length. In comparison, the set-associative arrays could be considered as a more flexible alternative that can be effectively used for achieving all three objectives.

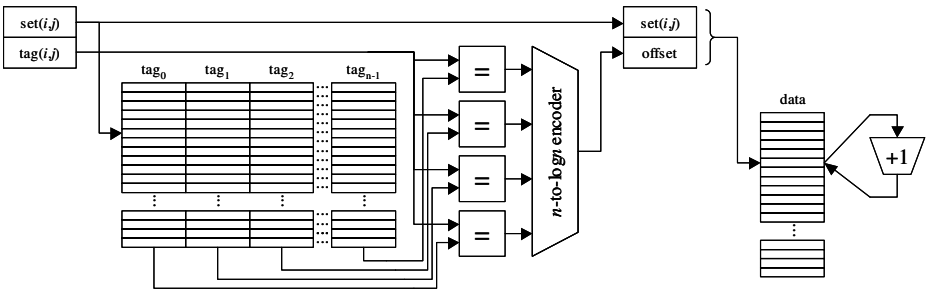


Fig. 2. Structure of the Co-occurrence Matrix Computation Unit

The internal structure of the CMCU is illustrated in Fig. 2. Every CMCU is implemented by means of an n -way set-associative array of N_c cells and auxiliary circuitry, which include n comparators, an n -to- $\log_2 n$ priority encoder and an adder. The set-associative arrays can be utilized for efficient storage and retrieval of sparse matrices, ensuring a throughput of one access per cycle with a latency of four cycles. An n -way set-associative array consists of n independent tag arrays ($\text{tag}_{s_0} - \text{tag}_{s_{n-1}}$). The tag-arrays are implemented in the distributed RAM of the FPGA and each of them consists of $N_c n$ cells. The set-associative array uniquely maps an input pair of 5-bit gray-level intensities (i, j) into an address of the N_c -cell data array. The data arrays are implemented using FPGA Block RAMs, each of which can hold up to 256 co-occurrence matrix elements. The data array cells contain the number of occurrences of the respective (i, j) pairs.

The circuit is implemented as a four-stage pipeline. In the first two cycles the circuit reads an input pair of gray-level intensities (i, j) and maps it to the address of the BlockRAM cell that stores $C_{d,\theta}(i, j)$. In the next two cycles the value of $C_{d,\theta}(i, j)$ is retrieved from the *data* array and incremented by one. The necessary forwarding

circuits are implemented, ensuring a stall free operation of the pipeline regardless of the input data, thus guaranteeing a throughput of one update operation per cycle with a latency of four cycles. After all input pairs (i,j) have been read and the corresponding cells have been updated, the unit outputs the computed GLCM.

3.2 Vector Calculation Units

The Vector Calculation Unit receives a GLCM computed by a CMCU and calculates the $\bar{V} = [V_1, V_2, V_3, V_4, V_5]$ (Eqs. 6-10) vector. The resulting vector is written to a bank of the on-card memory through the corresponding memory controller. The calculation of V_1 to V_4 is implemented in four independent pipelined circuits. The pipeline stages for each circuit are a preprocessing stage, calculation stages, a postprocessing stage and an accumulation stage. The preprocessing and postprocessing stages facilitate the operations needed to calculate the V_1 to V_5 from the lower triangular representation of the GLCM. In the preprocessing stage, the elements of the diagonal of the GLCM are doubled and in the postprocessing stage the intermediate results of the computation are multiplied by two for all elements that do not belong to the diagonal. The calculation stages involve LUT access or arithmetic operations such as addition, multiplication and subtraction. The output of each postprocessing stage is accumulated in a register during the accumulation stage. The intermediate results of each operation are not truncated or rounded, ensuring a high accuracy of the final results. The width of the integers increases after each arithmetic operation. The values of i and j are 5 bits wide and their product $i:j$ is 10 bits wide. The value of $C_{d,\theta}(i,j)$ is 16 bits wide and the product $i:j \cdot C_{d,\theta}(i,j)$ is represented by 26 bits. The accumulators in the final stage of the computation are 64 bits wide.

The calculation of V_5 (Eq. 10) is implemented in two separate pipelined circuits. The first pipeline has two stages and uses a 64×16 -bit BlockRAM for the storage of $C_x(i)$. At the first stage, the previous value of $C_x(i)$ is read from the BlockRAM and at the second stage it is incremented by $C_{d,\theta}(i,j)$ and stored back to the memory. A forwarding circuit ensures correct operation of the pipeline without stalls. The second pipeline is activated when all values $C_{d,\theta}(i,j)$ have been read and $C_x(i)$ has been calculated. It consists of three stages. At the first stage, $C_x(i)$ is retrieved from the BlockRAM, at the second it is squared and at the third it is accumulated into a register. The value of the accumulator after all $C_x(i)$ have been processed is the correct value of V_5 .

Computation of the Logarithm. To support the calculation of V_4 (Eq. 9), we implemented a method for the efficient approximation of the base-2 logarithm of 16-bit integers. This method results in a 3-stage pipelined circuit that requires 123 slices (less than 0.7% of the total FPGA area) and achieves a maximum frequency of 121.5MHz. The stages of the circuit are:

1. The integer part of the logarithm $l_i = \lfloor \log_2 n \rfloor$ is calculated by means of a priority encoder. Its value is the position of the most significant bit of the input integer.

$$2^k \leq n < 2^{k+1} \Rightarrow k \leq \log_2 n < k+1 \Rightarrow l_i = k \quad (18)$$

2. The fractional part of the logarithm $l_f = \log_2 n - l_i$ is estimated from Eq. (19), as a linear approximation between the points $(2^k, k)$ and $(2^{k+1}, k+1)$. The value l_f can be

easily extracted from the binary representation of n , by removing its most significant bit and right shifting by k bits.

$$\frac{n-2^k}{2^{k+1}-2^k} = \frac{l_i+l_f-l_i}{k+1-k} \Rightarrow l_f = \frac{n}{2^k}-1 \tag{19}$$

3. A novel method has been devised to increase the accuracy of this linear approximation of the logarithm. The fractional part of the logarithm l_f is transformed by Eq. (20).

$$l'_f = \begin{cases} (1+a) \cdot l_f & \text{if } l_f \leq 1/2 \\ (1-a) \cdot l_f + a & \text{if } l_f > 1/2 \end{cases} \tag{20}$$

The optimal a is the one that minimizes the error E between $\log_2 n$ and the approximated logarithm ($l_i + l'_f$), where

$$E = \frac{1}{65535} \sum_{n=1}^{65535} \frac{|\log_2 n - (l_i + l'_f)|}{\log_2 n} \tag{21}$$

The error E as a function of a is illustrated in Fig. 3

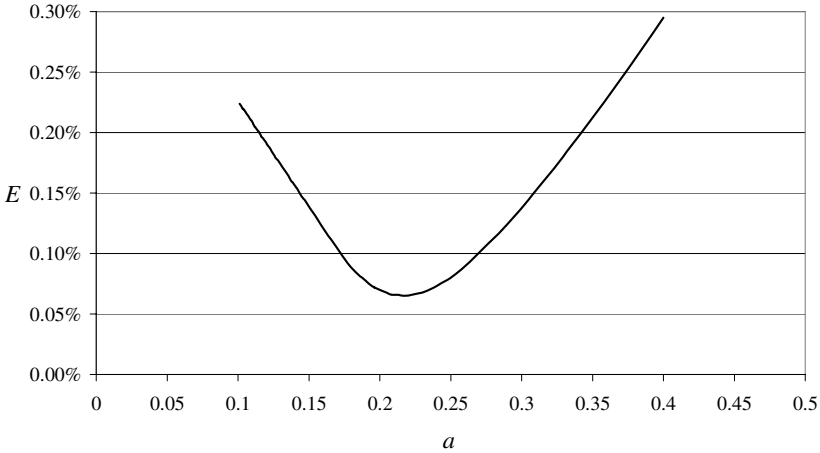


Fig. 3. Estimation of the error E for different values of a

Although the minimum error was achieved for $a=0.22$ ($E=0.07\%$), we selected $a=0.25$ ($E=0.08\%$) because it is implemented using simpler circuits (just a single shifter) and the error is only marginally higher.

4 Results

The proposed system was tested using natural raw images of 512×512 , 1024×1024 and 2048×2048 -pixel dimensions. The images were divided into blocks of 8×8 ,

16×16, 32×32, 64×64, 128×128 and 256×256-pixel dimensions and given as input to the system in order to evaluate its performance.

In the case of a 16×16-pixel or smaller input block, the triangular co-occurrence matrix for $N_g = 32$ is sparse, as the number of pixel pairs that can be considered for its computation, is smaller than the total number of co-occurrence matrix elements. Therefore, for input blocks of 8×8 and 16×16 pixels, N_c is set to the maximum possible value of 64 and 256 respectively.

In the case of a 32×32-pixel or a larger input block, the co-occurrence matrix is not considered sparse, as the number of all possible pixel pairs that take part in its computation is larger than the total number of its elements (i.e. 528). Therefore, N_c is set to 528.

By following a grid search approach for the determination of n , it was found that the sixteen-way set-associative arrays ($n = 16$) result in the optimal tradeoff between circuit complexity and time performance.

The proposed architecture, as implemented on the Xilinx Virtex-XCV2000E-6 FPGA, operates at 36.2MHz and 39.8MHz and utilizes 77% and 80% of the FPGA area for 8×8 and 16×16 input blocks respectively, by exploiting the sparseness of the co-occurrence matrices. The use of larger input blocks results in approximately the same operating frequency reaching 37.3MHz and an area utilization of 83%.

The image and block dimensions for which the proposed system achieves real-time performance are illustrated in Fig. 4.

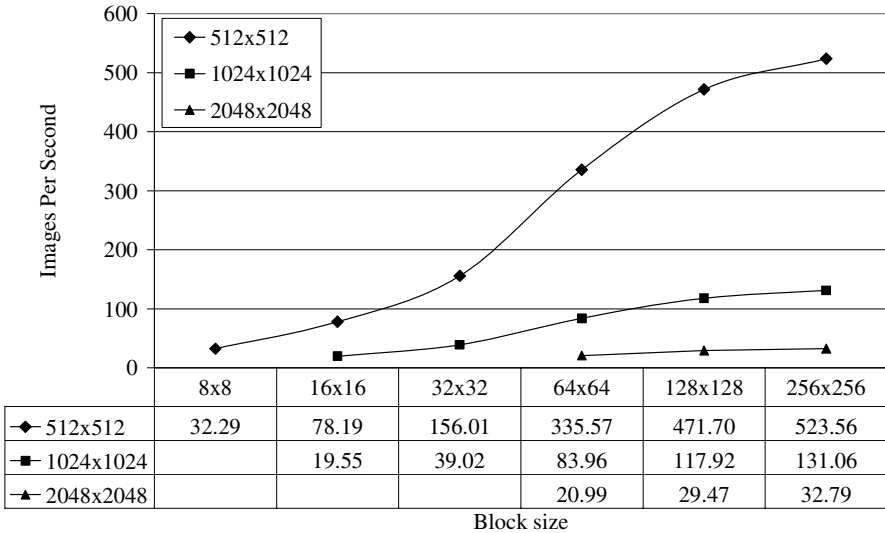


Fig. 4. Performance of the proposed system in images per second

The results show that as the dimensions of the block increases, the system performance is enhanced. It is worth noting that a real-time performance for video applications is reached using images of 2048×2048 pixels with a block size of 128×128 pixels or higher. Best time performance is achieved for 512×512-pixel images with a

block of 256×256 pixels, whereas marginal real-time performance is obtained for 1024×1024 -pixel images with a block of 16×16 pixels, and for 2048×2048 with a block of 64×64 pixels.

5 Conclusion

We presented a novel dedicated hardware system for real-time extraction of second-order statistical features from high-resolution images. It is based on a parallel FPGA architecture, enabling the concurrent calculation of sixteen gray-level co-occurrence matrices and four feature vectors. The input images are divided into blocks which are loaded to the RAM banks of the FPGA board. The FPGA processes each block and writes back four feature vectors. The performance of the proposed system increases as the image blocks become larger and the number of calculated vectors decreases.

The experimental results showed that the proposed system can be used for high resolution video applications which require real-time performance, the analysis of multiple video streams, and other demanding image analysis tasks.

The proposed system displays several advantages over the system presented in [2], which are summarized in the following:

- It calculates both the co-occurrence matrix and the features in a single FPGA core, whereas in [2] they are calculated in two separate cores. This avoids the overhead introduced by reprogramming each core onto the FPGA.
- It is capable of producing multiple feature vectors for each image, whereas in [2] a single feature vector is produced for each image. This facilitates in more accurate localization of texture within the image.
- It uses 25 bits per pixel for the representation of the input images, whereas in [2] the input images are represented using 5 bits per pixel. This allows a read rate of 25 bits per clock cycle from each memory bank, which results in a 5 times larger input bandwidth.
- It uses set-associative arrays for the sparse representation of the co-occurrence matrices, which enable the inclusion of four vector calculation units in a single core.

The results advocate the feasibility of real-time feature extraction from high-resolution images, using an efficient hardware implementation.

Future perspectives of this work include:

- The implementation of more second-order statistical features in a single FPGA core.
- The implementation of Color Wavelet Covariance (CWC) features [1] or other features based on co-occurrence matrices.
- The classification of feature vectors in hardware.

Acknowledgement

This research was funded by the Operational Program for Education and Vocational Training (EPEAEK II) under the framework of the project “Pythagoras - Support of

University Research Groups” co-funded by 75% from the European Social Fund and by 25% from national funds.

References

1. Karkanis, S.A., Iakovidis, D.K., Maroulis, D.E., Karras, D.A., Tzivras, M.: Computer Aided Tumor Detection in Endoscopic Video Using Color Wavelet Features. *IEEE Trans. Inf. Technol. Biomed.* 7 (2003) 141-152
2. Tahir, M.A., Bouridane, A., Kurugollu, F.: An FPGA Based Coprocessor for GLCM and Haralick Texture Features and their Application in Prostate Cancer Classification. *Anal. Int. Circ. Signal Process.* 43 (2005) 205-215
3. Baraldi, A., and Parmiggiani, F.: An Investigation of the Textural Characteristics Associated with Gray Level Cooccurrence Matrix Statistical Parameters. *IEEE Trans. Geosc. Rem. Sens.* 33 (2) (1995) 293-304
4. Shiranita, K., Miyajima, T., Takiyama, R.: Determination of Meat Quality by Texture Analysis. *Patt. Rec. Lett.* 19 (1998) 1319-1324
5. Iivarinen, J., Heikkinen, K., Rauhamaa, J., Vuorimaa, P., Visa, A.: A Defect Detection Scheme for Web Surface Inspection. *Int. J. Pat. Rec. Artif. Intell.* (2000) 735-755
6. Haralick, R.M., Shanmugam, K., Dinstein, I.: Textural Features for Image Classification. *IEEE Trans. Syst. Man Cybern.* 3 (1973) 610-621
7. Iakovidis, D.K., Maroulis, D.E., Karkanis, S.A., Flaounas, I.N.: Color Texture Recognition in Video Sequences Using Wavelet Covariance Features and Support Vector Machines. *Proc. 29th EUROMICRO*, Sept. 2003, Antalya, Turkey, pp. 199-204
8. Wei, C.-H., Li, C.-T., Wilson, R.: A Content-Based Approach to Medical Image Database Retrieval, in *Database Modeling for Industrial Data Management: Emerging Technologies and Applications*. ed. by Z. Ma, Idea Group Publishing, 2005
9. York, T.A.: Survey of Field Programmable Logic Devices. *Microprocessors and Microsystems.* 17 (7) (1993) 371-381
10. Ba, M., Degrugillier, D., Berrou, C.: Digital VLSI Using Parallel Architecture for Co-occurrence Matrix Determination. *Proc. Int. Conf. on Acoustics, Speech, and Signal Proc.*, 1989, Vol. 4, pp. 2556-2559
11. Heikkinen, K., Vuorimaa, P.: Computation of Two Texture Features in Hardware. *Proc. 10th Int. Conf. Image Analysis and Processing*, Sept. 1999, Venice, Italy, pp. 125-129
12. Theodoridis, S., Koutroumbas, K.: *Pattern Recognition*, Academic Press, San Diego (1999)
13. Haralick R.M.: Texture Measures for Carpet Wear Assessment, *IEEE Trans. Pattern Analysis and Machine Intelligence*, 10 (1) (1988) 92-104
14. Hennesy J.L., Patterson D.A.: *Computer Architecture, A Quantitative Approach*, Morgan Kaufmann, 2002

Greyscale Image Interpolation Using Mathematical Morphology

Alessandro Ledda¹, Hiêp Q. Luong¹, Wilfried Philips¹,
Valérie De Witte², and Etienne E. Kerre²

¹ Ghent University, TELIN-IPI, St. Pietersnieuwstraat 41, B-9000 Gent, Belgium
ledda@telin.UGent.be

<http://telin.ugent.be/~ledda/>

² Ghent University, Department of Applied Mathematics & Computer Science,
Krijgslaan 281, S9, B-9000 Gent, Belgium

Abstract. When magnifying a bitmapped image, we want to increase the number of pixels it covers, allowing for finer details in the image, which are not visible in the original image. Simple interpolation techniques are not suitable because they introduce jagged edges, also called “jaggies”.

Earlier we proposed the “mmINT” magnification method (for integer scaling factors), which avoids jaggies. It is based on mathematical morphology. The algorithm detects jaggies in magnified binary images (using pixel replication) and removes them, making the edges smoother. This is done by replacing the value of specific pixels.

In this paper, we extend the binary mmINT to greyscale images. The pixels are locally binarized so that the same morphological techniques can be applied as for mmINT. We take care of the more difficult replacement of pixel values, because several grey values can be part of a jaggy. We then discuss the visual results of the new greyscale method.

1 Introduction

A bitmap, or raster graphic, consists of *pixels*, aligned on a grid. The scene in the image is described in terms of those pixels’ values. In order to magnify a bitmap image, the same scene must be represented using more pixels, i.e., image interpolation is needed.

The simplest interpolation method is *pixel replication* or *nearest neighbour interpolation*; pixel values in the enlarged image are copied from the pixels at the corresponding position in the original image, or – if that position does not correspond to a pixel centre – from the pixels nearest to that position. Fig. 1 shows an example for 4 times magnification. The result contains unwanted jagged edges, called *jaggies*. For non-integer scaling, other artefacts such as aliasing are even more undesirable.

Other linear (non-adaptive) techniques are the *bilinear* and *bicubic interpolation* [1]. Here, the new pixel values are computed as the (weighted) mean of the 4 and 16 closest neighbours, respectively. Other non-adaptive methods use higher order (piecewise) polynomials, B-splines, truncated or windowed sinc functions,

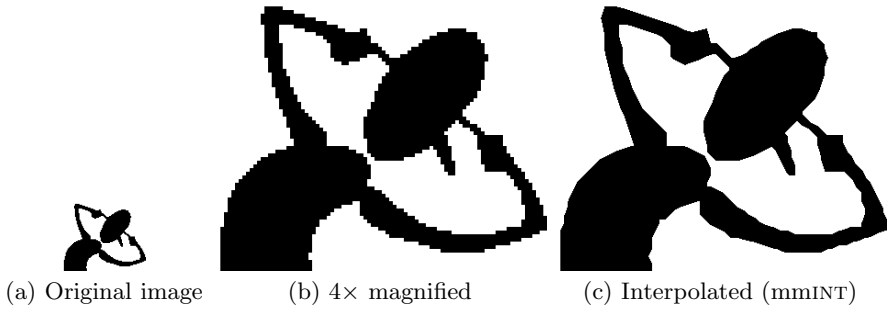


Fig. 1. Pixel replication (b) creates “jaggies”

etc. They create a greyscale image even for a binary input and most of them introduce additional artefacts, e.g., blurring and/or ringing.

Adaptive or non-linear interpolation methods incorporate prior knowledge about images to achieve better interpolation results. *Edge-based techniques* follow the principle that no interpolation across the edges in the image is allowed or that interpolation has to be performed along the edges. Examples of these techniques are EDI [2], NEDI [3] and Aqua [4]. *Restoration methods* use regularization methods or smoothing to limit interpolation artefacts. Some restoration methods use PDE-based regularization [5], isophote smoothing [6], level curve mapping [7] and mathematical morphology [8]. Our proposed method mmINT(g) (see fig. 1(c)) also belongs to the class of restoration interpolation techniques.

Some other adaptive methods exploit the self-similarity property of an image, e.g., methods based on iterated function systems [9]. Example-based approaches are yet another class of adaptive interpolation methods. They map blocks of the low-resolution image into pre-defined interpolated patches [10, 11]. Adaptive methods still suffer from artefacts: their results often look segmented, bear important visual degradation in fine textured areas or random pixels are created in smooth areas.

In earlier work [12], we presented a non-linear interpolation technique for *binary input images* based on mathematical morphology, *mmINT* (*Mathematical Morphological INTerpolation*). The technique performed quite well and had the advantage of producing a binary output image. The basic idea of the method is to detect jaggies from a pixel replicated image and then iteratively correct them by inserting “corner pixels” in or removing them from those jagged edges, while taking care not to distort real corners.

In this paper we present a greyscale extension of mmINT. The new method is called mmINTg. The basic idea is to locally binarize the greyscale input image, in order to detect the corners of the jagged edges, and then change their pixel values so the edges become smooth, but still remain sharp and not blurred. mmINTg is backward compatible with mmINT, in the sense that the interpolation result of mmINTg on a binary input image is binary too.

The next section gives an introduction to mathematical morphology and the morphological hit-miss transform used in our algorithm. Section 3 summarizes

the principles of mmINT [12], while section 4 describes the changes made for the greyscale extension (mmINTg). Then we discuss the visual results of mmINTg and we will conclude that this method is good at interpolating line drawings.

2 Theoretical Background

2.1 Morphological Operators

Mathematical morphology [13, 14, 15], originally developed as a theory for binary images, is a framework for image processing based on set theory. Morphological image processing can simplify images, preserving objects' essential shape characteristics, while eliminating irrelevant objects or smoothing their border.

Binary mathematical morphology is based on two basic operators: *dilation* and *erosion*. The dilation and erosion of a set A by a structuring element (short: strel) B are defined as:¹

$$\begin{aligned} \text{Dilation : } A \oplus B &= \bigcup_{\mathbf{b} \in B} T_{\mathbf{b}}(A) \\ \text{Erosion : } A \ominus B &= \bigcap_{\mathbf{b} \in B} T_{-\mathbf{b}}(A) , \end{aligned} \quad (1)$$

with $T_{\mathbf{b}}(A)$ the translation of image A over the vector \mathbf{b} . Computing $A \oplus B$ and $A \ominus B$ amounts to sliding the strel over the input image and computing for each position an output pixel based on the contents of the strel and the contents of the corresponding part of A .

In general, a dilation results in an enlarged version of the input object. The net effect of an erosion is to shrink or erode the input object.

The structuring element B can be of any size or shape and is chosen depending on the image and the application. The origin of the strel is also important, as it states how the strel is positioned relative to the examined pixel.

2.2 The Hit-Miss Transform

The *hit-miss transform* “ \otimes ” is a morphological operator used extensively in our algorithm. It is defined in terms of two disjoint structuring elements: one for the erosion of object pixels (B), and one for the erosion of background pixels (C). Its definition is:

$$A \otimes (B, C) = (A \ominus B) \cap (A^c \ominus C) , \quad (2)$$

with A^c the complement set of A .

With the hit-miss transform it is possible to detect specific shapes in the image. We will use it to detect corners. To detect the upper-left corners of an object, we use the structuring elements (a) and (b) of fig. 2. An alternative for strel C for the corner detection is the use of the structuring element (c) in fig. 2.

The result of the hit-miss transform is an image in which the foreground pixels indicate the position of the upper-left corners. Alternatively, the output can also

¹ A is the set of the coordinates of the foreground pixels (value 1) in an image. In the remainder of this paper, we will often simply refer to A as “the (binary) image”.

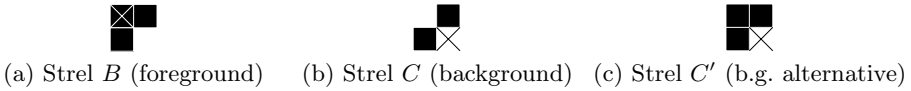


Fig. 2. Upper-left corner detection with the hit-miss transform. Specific strels are used. The black squares are pixels of the strel; the cross is the origin of the strel.

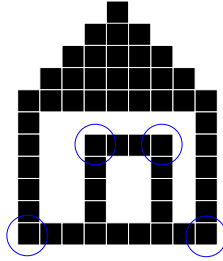


Fig. 3. The difference between “jagged corners” and “real corners” (encircled)

be viewed as a set of corner coordinates. We further refer to this set as a *corner map*. In our method we apply the hit-miss transform several times, with rotated versions of the structuring elements in fig. 2 in order to produce four corner maps (upper-left, upper-right, lower-left and lower-right).

3 Binary Interpolation Scheme: mmINT

In this section we summarize the mmINT method presented in [12]. Its purpose is to remove the jagged edges from a pixel replicated image, by swapping specific pixels from background to the foreground and vice versa. We consider the most frequent colour in the image to be the background.

Different steps can be distinguished in the algorithm:

1. **Pixel replication:** First the image is pixel-replicated by an integer factor M . The resulting image contains strong staircase patterns because of the pixel replication (see for example fig. 1).
2. **Corner detection:** Using a combination of hit-miss transforms, the algorithm determines the positions of corners, both real and false (due to jaggies) in the image.
3. **Corner validation:** Some corners found in the preceding step are *real corners*, which have to be retained in the interpolated image. For example, the corners of the door and walls in fig. 3 are real corners. The corners detected on the roof are *jagged corners*, because the ideal roof is a diagonal line. The aim of corner validation is to distinguish false corners from real ones.
4. **Pixel swapping (interpolation):** We swap the colours of pixels classified as false corners, and the colours of some of their neighbours.

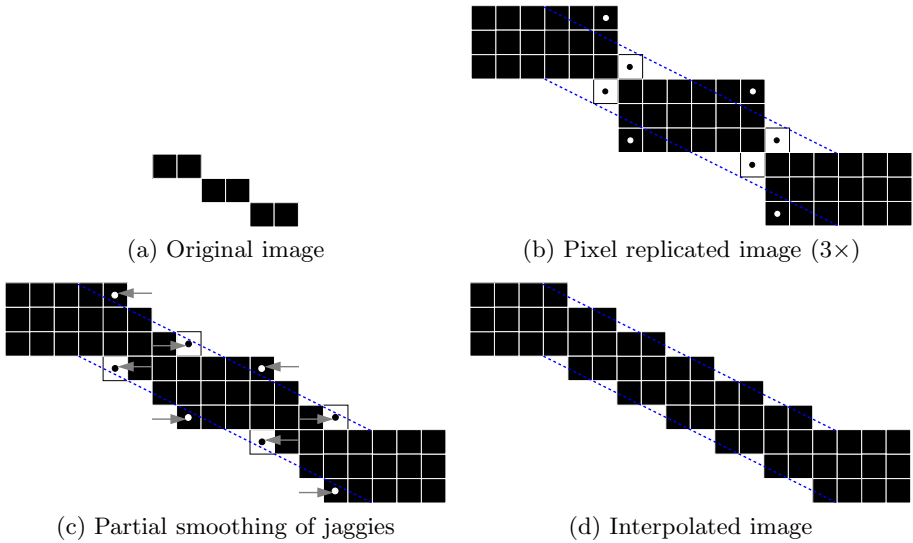


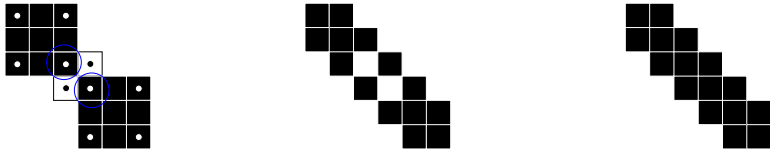
Fig. 4. The jagged edges have to be removed, by replacing the values of specific pixels. The dotted lines show the orientation of the original line (a). The dots show the pixels that will change value after interpolation.

The above operations (except the first one) are repeated iteratively, each iteration operating on the output image from the preceding iteration.

In order to illustrate the details of our method, we consider the case of enlarging an object consisting of a thin line of foreground pixels (see fig. 4). The result after pixel replication is shown in fig. 4(b) and is clearly jagged. The dotted lines show the ideal boundaries of the magnified line. The ideal solution would be to replace all background (white) pixels between the dotted lines with foreground (black) pixels and to replace all foreground pixels outside the dotted lines with background pixels. The iterative procedure aims to do just that.

As illustrated in fig. 4(b), the jaggies can be removed by altering the pixel values at the locations of object and background corners. The positions of these corners can be located with the morphological hit-miss transform as explained before, using the structuring elements from fig. 2. We not only look for corners of the objects, but also for corners in the background. This way, we will have 8 corner maps (4 *object* corner maps and 4 *background* corner maps).

If we use the same structuring elements for the detection of object corners and background corners, then artefacts will occur at lines with barely touching pixels (see fig. 5 ($M = 3$)), i.e., pixels with the same value that are only connected by 8-connectivity. Two object corners and two background corners are found at such 8-connectivity. If all those corner pixels change value, then holes are introduced (fig. 5(b)). To avoid these artefacts, the structuring element shown in fig. 2(b) is replaced by the one shown in fig. 2(c): for the detection of an upper-left corner, not only the pixel values to the left and above the current pixel are investigated,



(a) Before pixel swapping (b) Without compensation (c) With more strict strels

Fig. 5. (a) Barely touching pixels (encircled) could give (b) artefacts after interpolation. (c) The use of more strict strels for the detection of object corners prevents this.



(a) Complements across the edge (b) Complements along the edge

Fig. 6. Complementary corners: the background corner (black dot) has different complementary corners (white dots) at specific relative coordinates. Magnification $M = 3$.

but also the neighbouring pixel at the upper-left. This new strel is only used for the detection of a foreground corner, because otherwise no interpolation will take place at all at lines with barely touching pixels.

In the corner validation step, we distinguish between real and jagged corners by searching for one or more *complementary corner pixels* in a local neighbourhood, and this for every detected corner pixel, in each of the 8 corner maps. A complementary corner of an object corner c_o is a corner in the background (or vice versa) that lies at specific relative coordinates w.r.t. c_o . Fig. 6(a) shows a pixel-replicated line (magnification $M = 3$) with a background corner and two complementary corners across the edge at a distance of M pixels. Complements along the edge (fig. 6(b)) are located at $M - 1$ pixels in one direction and $1 + (\theta - 1)(M - 1)$ in the other direction (with θ the iteration step). The existence of a complementary corner indicates the presence of a jagged edge, and thus a corner with at least one complementary corner will not be removed from the corner map.

In the pixel swapping step, we change the value (0 or 1) of the pixels (and surrounding neighbours) that are detected as corners of a jagged edge. Which pixels exactly need to change, is defined by the n times dilation of the corner maps with the structuring element B , with B shown in fig. 8(a) for the upper-left corner map. $n = \lfloor \frac{M}{2} \rfloor - 1$ for odd magnification M ; for even M , $n = \lfloor \frac{M}{2} \rfloor - 1$ for the background and $n = \lfloor \frac{M}{2} \rfloor - 2$ for the objects in the odd iterations, and the situation is reversed in the even iterations. If $n < 0$, then no pixels are swapped.

At this point, lines with angles other than 0° , $\pm 45^\circ$ or 90° are not yet completely smooth. In our example, fig. 4(c) shows the result of the first iteration step of our method, while we want to obtain fig. 4(d). Therefore the algorithm

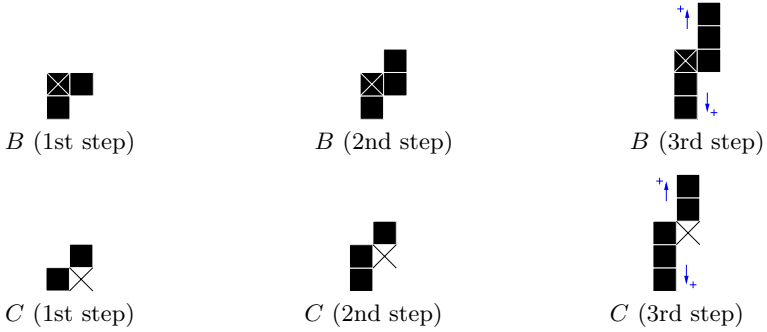


Fig. 7. The hit-miss structuring elements are different for every iteration step

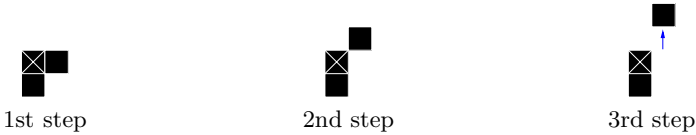


Fig. 8. The interpolation structuring element is different for every iteration step

is repeated until all appropriate changes have been made (on average 15 times), using different (and larger) structuring elements in successive iteration steps for the hit-miss transform (see fig. 7) and pixel swapping step (see fig. 8). Notice that the structuring elements are less symmetric for $\theta > 1$, so the number of corner maps doubles because we also have to look for corners using the mirrored strels, besides rotated versions [12].

The interpolation result of fig. 1(a) using *mmINT* is shown in fig. 1(c).

4 Extension to Greyscale: *mmINTg*

In this paper we introduce the greyscale extension of *mmINT*. In the binary case, only two possibilities exist: a pixel is part of the fore- or the background. Also, when we look for jagged edges, the result is a detection of a corner or no corner.

The greyscale case is more complicated: the classification of a pixel to the foreground or the background is not straightforward, which makes the detection of corners using the binary hit-miss transform more difficult. We therefore adapt the corner detection step (see section 4.1). For this purpose the pixels are locally binarized, before applying the hit-miss operation. A majority ordering is used for the classification of a pixel as a foreground or a background corner.

In the pixel swapping step (see section 4.2), the values of the neighbouring pixels are taken into account to calculate the interpolated pixel value. This algorithm for greyscale interpolation is called *mmINTg*.

We will now discuss the new corner detection and pixel swapping algorithm.

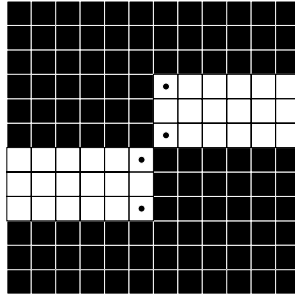


Fig. 9. If white is considered background, then no black corners will be detected, which results in a “real” edge

4.1 New Corner Detection Method for Step 2

Binarization. In order to apply the binary hit-miss transform to detect corners, we binarize the pixel values. A greyscale hit-miss transform would also be possible [16, 14], but then the result is still a greyscale image, while we wish for a binary result (i.e., corner or not). The binarization is done locally and only the pixels that are then covered by the structuring elements are taken into account (see fig. 2 for the strels used in the first iteration step). The threshold value $T(x, y)$ is defined by:

$$T(x, y) = \begin{cases} \frac{1}{2}(m + M) + \alpha, & \text{if } I(x, y) \geq \frac{1}{2}(m + M) \\ \frac{1}{2}(m + M) - \alpha, & \text{if } I(x, y) < \frac{1}{2}(m + M) \end{cases}, \quad (3)$$

with m and M respectively the minimum and maximum value of the set of pixels defined by the structuring elements, $I(x, y)$ is the grey value of the currently checked pixel, and α is a threshold for classifying more neighbouring pixels into a class different to the one of the current pixel. We have experimentally found that $\alpha = (M - m)/10$ is a good choice. The current pixel is always given the binary value 1, the value of the other considered pixels depends on their classification w.r.t. the current pixel.

Majority Ordering. In section 3 we mentioned that different strels are used for foreground and background corner detection. Also, the corner validation step looks at complementary corners, which are part of the opposite class as the pixel under investigation. This means that we need to classify the pixels as either possible object corner or background corner.

We utilize the *majority sorting scheme* (MSS) [17, 18] to order the grey values locally in function of their presence in a local window. If the grey value of the currently investigated pixel appears less than the grey values of the other pixels covered by the strels, then the pixel is considered foreground and strels (a) and (c) from fig. 2 are used for the upper-left corner detection.

The area in which we calculate an ordering map with the MSS is an 11×11 window in the original low-resolution image, since this size gives satisfying

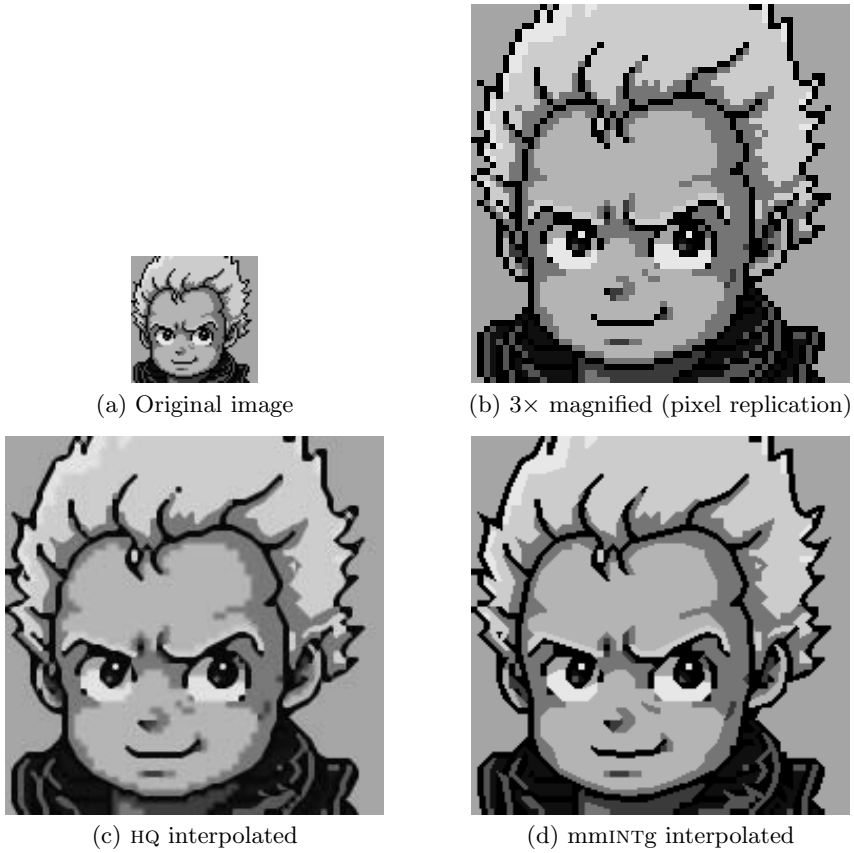


Fig. 10. The interpolation of a greyscale cartoon sprite

interpolation results. If the window is taken too small, then the ordering will be less accurate, because the probability to get the same counts for different grey values will become higher. If we take the window too large, then some pixels might not be interpolated. For example, fig. 9 shows a white line on a black background. We expect this line to be interpolated, but if in a larger window the white pixels are in the majority, then white is considered as background. In this case, no interpolation will occur, since different structuring elements are used for object and background corner detection. There will only be detection of white corners, which will not pass the corner validation step, because there are no black corners detected in their neighbourhood.

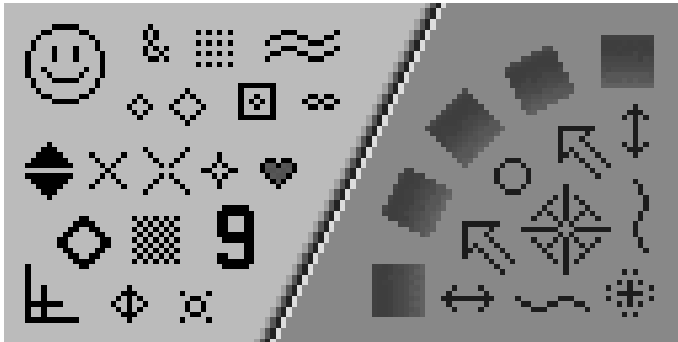
The majority ordering only has to be performed in the first iteration step. From then on we know whether the pixel belongs to the fore- or background.

4.2 New Interpolation Method for Step 4

In the binary case, the values of the jagged corner pixels and surroundings are replaced by the opposite colour, i.e., black becomes white and white becomes



(a) Original image



(b) 3x magnified (pixel replication)



(c) HQ interpolated



(d) mmINTg interpolated

Fig. 11. The interpolation of a greyscale line graphic

black. With greyscale images, we cannot simply swap the pixel values to the opposite grey value. The grey values of the surrounding pixels must be taken into account, so that the transition between grey values does not occur abruptly. The new pixel value is the average grey value of the pixels that are covered by the background hit-miss structuring element when positioned at the corner pixel (see fig. 2(b) for the strel used in the first iteration step). The resulting value is thus defined in function of the surrounding values of the other class. For a binary image, the effect is the same as with mmINT, since the average is taken of pixels of the same colour, a colour that is opposite to that of the current pixel. As a consequence, no blurring occurs.

5 Results

mmINT is a technique that is very good at interpolating line art images, like logos, cartoons, maps, etc. Our proposed greyscale extension, mmINTg, also performs very well on this kind of images. When binary images are processed, the same results can be expected with mmINTg as with mmINT. When the MSS locally produces the opposite ordering as on the entire image (which is done for mmINT to define the background colour), this will lead to slightly different results. Situations like the ones in fig. 9 will be tackled by mmINTg.

The figures 10 and 11 show images with clear sharp edges. We compare our greyscale method with HQ, a technique that is very competitive with mmINT when interpolating binary images [10,12]. In the first figure, we notice that mmINTg interpolates the lines better and the result is less blurred. The results

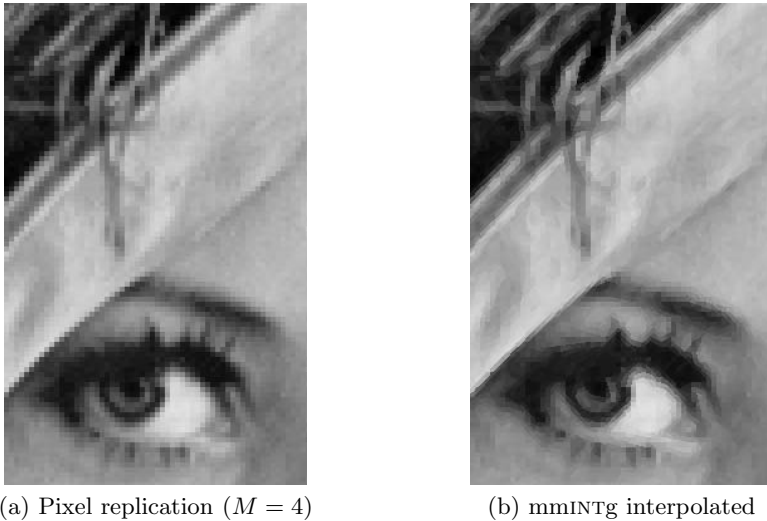


Fig. 12. The interpolation with mmINTg of a real life scene (a cut-out of the Lena image)

shown in fig. 11 look very similar, except that the rotated squares look blurred with HQ. Also, HQ introduces too much unnecessary colours in the arrows at the right, i.e., when 8-connectivity lines are present.

Fig. 12 shows a real life scene containing a lot of texture and grey value variation. mmINTg produces quite good results, but because the edges in the image are less sharply defined and more grey values are involved, the interpolation looks segmented, i.e., the grey value change after interpolation is too abrupt.

6 Conclusions

We presented a greyscale solution for the binary morphological interpolation technique mmINT. A temporary local binarization of the grey values, combined with a local majority ordering, makes it possible to perform the morphological hit-miss transform on the image. The grey values of the pixels that need to change value, are defined in terms of neighbouring pixel values.

The visual quality of the new mmINTg is very good for cartoon sprites and line graphics. Its results are in most cases visually better than those of HQ.

References

1. Lehmann, T., Gönner, C., Spitzer, K.: Survey: Interpolations Methods In Medical Image Processing. *IEEE Transactions on Medical Imaging* **18** (1999) 1049–1075
2. Allebach, J., Wong, P.: Edge-directed interpolation. In: *Proceedings of the IEEE International Conference on Image Processing ICIP '96*. Volume 3., Lausanne, Switzerland (1996) 707–710
3. Li, X., Orchard, M.: New Edge-Directed Interpolation. *IEEE Transactions on Image Processing* **10** (2001) 1521–1527
4. Muresan, D., Parks, T.: Adaptively quadratic (AQua) image interpolation. *IEEE Transactions on Image Processing* **13** (2004) 690–698
5. Tschumperlé, D.: PDE's Based Regularization of Multivalued Images and Applications. PhD thesis, Université de Nice — Sophia Antipolis, Nice, France (2002)
6. Morse, B., Schwartzwald, D.: Isophote-Based Interpolation. In: *Proceedings of the IEEE International Conference on Image Processing ICIP '98*, Chicago, Illinois, USA (1998) 227–231
7. Luong, H., De Smet, P., Philips, W.: Image Interpolation using Constrained Adaptive Contrast Enhancement Techniques. In: *Proceedings of the IEEE International Conference on Image Processing ICIP '05*, Genova, Italy (2005) 998–1001
8. Albiol, A., Serra, J.: Morphological Image Enlargements. *Journal of Visual Communication and Image Representation* **8** (1997) 367–383
9. Honda, H., Haseyama, M., Kitajima, H.: Fractal Interpolation for Natural Images. In: *Proceedings of the IEEE International Conference on Image Processing ICIP '99*. Volume 3., Kobe, Japan (1999) 657–661
10. Stepin, M.: hq3x Magnification Filter. <http://www.hiend3d.com/hq3x.html> (2003)
11. Freeman, W., Jones, T., Pasztor, E.: Example-Based Super-Resolution. *IEEE Computer Graphics and Applications* **22** (2002) 56–65

12. Ledda, A., Luong, H., Philips, W., De Witte, V., Kerre, E.: Image Interpolation using Mathematical Morphology. In: Proceedings of 2nd IEEE International Conference on Document Image Analysis for Libraries (DIAL'06), Lyon, France (2006) 358–367
13. Serra, J.: Image Analysis and Mathematical Morphology. Volume 1. Academic Press, New York (1982)
14. Soille, P.: Morphological Image Analysis: Principles and Applications. 2nd edn. Springer-Verlag (2003)
15. Haralick, R., Shapiro, L.: 5. In: Computer and Robot Vision. Volume 1. Addison-Wesley (1992)
16. Ronse, C.: A Lattice-Theoretical Morphological View on Template Extraction in Images. *Journal of Visual Communication and Image Representation* **7** (1996) 273–295
17. Ledda, A., Philips, W.: Majority Ordering for Colour Mathematical Morphology. In: Proceedings of the XIIIth European Signal Processing Conference, Antalya, Turkey (2005)
18. Ledda, A., Philips, W.: Majority Ordering and the Morphological Pattern Spectrum. In: Proceedings of ACIVS. Volume 3708 of Lecture Notes in Computer Science., Antwerp, Belgium, Springer (2005) 356–363

Dilation Matrices for Nonseparable Bidimensional Wavelets

Ana Ruedin

Departamento de Computación, Facultad de Ciencias Exactas y Naturales,
Universidad de Buenos Aires
Ciudad Universitaria, Pab. I. CP 1428, Ciudad de Buenos Aires
`anita@dc.uba.ar`

Abstract. For nonseparable bidimensional wavelet transforms, the choice of the dilation matrix is all-important, since it governs the downsampling and upsampling steps, determines the cosets that give the positions of the filters, and defines the elementary set that gives a tessellation of the plane. We introduce nonseparable bidimensional wavelets, and give formulae for the analysis and synthesis of images. We analyze several dilation matrices, and show how the wavelet transform operates visually. We also show some distortions produced by some of these matrices. We show that the requirement of their eigenvalues being greater than 1 in absolute value is not enough to guarantee their suitability for image processing applications, and discuss other conditions.

Keywords: quincunx, nonseparable, dilation, wavelet.

1 Introduction

In the last 20 years, wavelet transforms have acquired great importance. They have proved efficient in many domains, and the number of their applications is continually increasing. As a tool in image processing, wavelet transforms are being applied to image compression (they are the basis of the standard JPEG 2000 [1]), image denoising, texture analysis and pattern recognition.

Wavelets functions are bases which have excellent time (or space) localization, and good frequency localization. In one dimension they are dilations (changes in scale in powers of 2) and integer translations of one same function $\Psi(x)$ called wavelet. Representing a signal in these bases, we obtain a decomposition of the signal in sums of details of different resolutions, plus a coarse approximation of the signal.

The wavelet transform can be calculated by means of the fast wavelet transform, through convolutions of the signal with two filters, one lowpass filter and one highpass filter, followed by a decimation by 2, i.e. only the even coefficients are retained. Because of the properties of the wavelet, this operation is reversible.

To process an image, the traditional manner is to apply twice a one-dimensional wavelet transform, first by rows and then by columns. This is called the separable

bidimensional transform. It originates 3 different detail coefficients, in the horizontal, vertical and diagonal directions, which does not correspond to our visual system. Seeking a more isotropic treatment of the details of an image, more generalized wavelets were investigated, and truly bidimensional wavelets were found. These functions, defined on the real plane, cannot be factorized into a product of functions of one argument. They are called nonseparable wavelets.

Since bidimensional nonseparable wavelets are defined on the plane, their dilation factor is a 2×2 matrix, called the dilation matrix. It can be equal to twice the identity matrix, or can have other values. Its elements must be integers, and it must be an expansion of the plane. For this, the requirement generally used is that its eigenvalues be > 1 (in absolute value). Some authors ([2], [3]), require instead that the singular values be > 1 .

If $|D| = 2$, the number of wavelets associated to one scaling function is 1. That is, at each scale of the wavelet transform, we will have only one kind of detail coefficients.

The choice of the dilation matrix is all-important. It governs the decimation (or downsampling) and the upsampling steps of the wavelet transform or antitransform. It determines a grid that defines the positions of the filters. The orthogonality condition of the filter bank, the degree of polynomial approximation of the scaling function, and the behaviour of the filters (as good lowpass or highpass filters) can be written in terms of the filters and their positions, which depend on the dilation matrix (see [4], [2], [5] [6] [7]). Finally, in order to prove the Holder continuity of the wavelets, a tessellation of the plane has to be made with shifts of an elementary set [8], [2]. This elementary set is completely determined by the dilation matrix.

Before applying these nonseparable wavelet transforms, it is imperative to know what visual effects will result.

We shall analyze four different dilation matrices, having determinant equal to 2 (in absolute value). They are D_i , for $i = 1 \dots 4$; their values are given in Section 3.

Kovacevic and Vetterli [4], Cohen and Daubechies [2], found examples of nonseparable wavelets with matrices D_1 and D_2 .

Ayache [9], Kovacevic and Vetterli [10], He and Lai [11] and Faugère et. al. [12] gave examples of nonseparable wavelets with dilation matrix $D = 2I$.

Belogay and Wang [13] found examples of orthogonal, nonseparable wavelets with dilation matrix D_4 .

In previous papers were constructed examples of nonseparable multiwavelets for matrices D_1 and D_2 ([14], [15], [7]).

In [16], dilation matrices are analyzed for cubic lattices.

In this paper, we analyze different possible dilation matrices for the plane. We introduce nonseparable bidimensional wavelets, and give formulae for the analysis and synthesis of images. We show pictures of wavelet processing for D_1 . We show that the requirement (on the dilation matrices) of their eigenvalues being greater than 1 in absolute value is not enough to guarantee their suitability for image processing applications, and discuss this condition versus the condition on the singular values.

2 The Nonseparable Bidimensional Scaling Function and Wavelet

Wavelets are associated to a scaling function. A one dimensional scaling function $\Phi(x)$ verifies a dilation or refinement equation

$$\Phi(x) = \sum_{k=0}^N h_k \Phi(2x - k).$$

This indicates that Φ is equal to a weighted sum of integer shifts of a compressed version of itself. To obtain a compressed version of Φ , we multiply the argument by a factor of 2.

For the nonseparable bidimensional case, a scaling function

$$\Phi : \mathfrak{R}^2 \rightarrow \mathfrak{R},$$

is a function that also verifies a dilation or refinement equation:

$$\Phi(x) = \sum_{k \in A \subset Z^2} h_k \Phi(D x - k), \tag{1}$$

where the values h_k correspond to a 2 dimensional filter, $k = (k_1, k_2) \in Z^2$, and D is a 2×2 matrix called dilation matrix. We multiply $x = (x_1, x_2) = [x_1 \ x_2]^T$ by the dilation matrix D , to obtain a condensed version of Φ .

By abuse of notation, $\Phi(D x - k)$ indicates that we apply Φ to the 2 components of vector

$$D x - k = \begin{bmatrix} d_{11} & d_{12} \\ d_{21} & d_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} - \begin{bmatrix} k_1 \\ k_2 \end{bmatrix}.$$

The approximation spaces V_j are generated by integer shifts of the scaling function Φ , or a contracted or dilated version of Φ :

$$V_j = \overline{\text{gen}\{\Phi(D^j x - k)\}_{k \in Z^2}}.$$

They are nested subspaces

$$\dots \subset V_{-2} \subset V_{-1} \subset V_0 \subset V_1 \subset \dots,$$

their union is dense in $L^2(\mathfrak{R})$, and their intersection is the null function.

In the following we shall only consider the cases in which the set

$$\{\Phi(x - k)\}_{k \in Z^2}$$

is orthonormal.

A nonseparable bidimensional wavelet verifies:

$$\begin{aligned} \Psi : \mathfrak{R}^2 &\rightarrow \mathfrak{R}, \\ \Psi(x) &= \sum_{k \in A \subset Z^2} g_k \Phi(D x - k), \end{aligned} \tag{2}$$

where g is a 2 dimensional filter.

Table 1. Properties of the dilation matrices

Dilation matrix	Determinant	eigenvalues	singular values
D_1	-2	$\lambda_1 = +\sqrt{2}$ $\lambda_2 = -\sqrt{2}$	$\sigma_1 = \sqrt{2}$ $\sigma_2 = \sqrt{2}$
D_2	2	$\lambda_1 = 1 + i$ $\lambda_2 = 1 - i$	$\sigma_1 = \sqrt{2}$ $\sigma_2 = \sqrt{2}$
D_3	2	$\lambda_1 = 2$ $\lambda_2 = 1$	$\sigma_1 = 2.28824561$ $\sigma_2 = 0.874032$
D_4	-2	$\lambda_1 = \sqrt{2}$ $\lambda_2 = -\sqrt{2}$	$\sigma_1 = 2$ $\sigma_2 = 1$

The detail subspace W_j is the orthogonal complement of V_j in V_{j+1} ,

$$V_{j+1} = V_j \oplus W_j,$$

and it is generated by shifts of the wavelet

$$W_j = \overline{\text{gen}\{\Psi(D^j x - k)\}_k}.$$

Remark 1. If the dilation matrix is $D = 2I$, and if there exist onedimensional filters f and ℓ such that $h_{i,j} = f_i f_j$ and $g_{i,j} = \ell_i \ell_j$, then h and g are separable filters, and we have a separable wavelet.

3 Dilation Matrices

As stated, we shall deal with dilation matrices having determinant equal to ± 2 . Their entries must be integers, and we require that their eigenvalues $|\lambda_k| > 1$; although we shall allow one of the eigenvalues to have norm 1. With these requirements, four different matrices can be constructed; others will be a permutation of these four.

We shall focus our analysis on

$$D_1 = \begin{bmatrix} 1 & 1 \\ 1 & -1 \end{bmatrix}, \quad D_2 = \begin{bmatrix} 1 & -1 \\ 1 & 1 \end{bmatrix}, \quad D_3 = \begin{bmatrix} 2 & 1 \\ 0 & 1 \end{bmatrix} \quad \text{and} \quad D_4 = \begin{bmatrix} 0 & 2 \\ 1 & 0 \end{bmatrix}.$$

D_1 is a reflection on axis $[1 \ \sqrt{2} - 1]^T$, followed by an expansion in $\sqrt{2}$.

D_2 is a rotation in $\frac{\pi}{4}$, followed by an expansion in $\sqrt{2}$. D_3 is a shear.

D_4 is an expansion in one direction, followed by a reflection on $x_2 = x_1$.

In table 1 are listed their eigenvalues and singular values.

4 Cosets

A dilation matrix D induces a decomposition of the set of all pairs of integers, into several cosets. The number of cosets is determined by the number of classes

of quotient Z^2/DZ^2 , which is equal to $|D| = |\det(D)|$. If all the entries of D are integers, then its determinant is also an integer.

In the cases we are concerned with, $|D|=2$, and we have 2 cosets Γ_0 and Γ_1 ,

$$Z^2 = \Gamma_0 \cup \Gamma_1 \ ; \ \Gamma_0 = \{DZ^2\} \ ; \ \Gamma_1 = \left\{DZ^2 + \begin{bmatrix} 1 \\ 0 \end{bmatrix}\right\}. \tag{3}$$

In the following, the cosets are drawn with crosses and circles.

For dilation matrices D_1, D_2 and D_3 , we have the following cosets

$$\begin{matrix} \text{O} \times \text{O} \times \text{O} \\ \times \text{O} \times \text{O} \times \\ \text{O} \times \text{O} \times \text{O} \end{matrix} \tag{4}$$

which are displayed as the black or white squares of a chessboard. They conform the quincunx grid.

For dilation matrix D_4 , we have a different grid. The 2 cosets are interleaved columns of Z^2 .

$$\begin{matrix} \times \text{O} \times \text{O} \times \\ \times \text{O} \times \text{O} \times \\ \times \text{O} \times \text{O} \times \end{matrix}$$

5 Up and Downsampling with D

When downloading an image from a database, first a small image with poor definition is shown (thumbnail), which enables the user to recognize the image promptly and cut down or finish the transmission. Then a progressive transmission is done. The thumbnail image may be obtained from the approximation coefficients of the wavelet transform, which includes many downsampling steps to reduce the size of the image. The process is useless if the thumbnail image is distorted.

Let us see the definition and the visual effects of downsampling and upsampling with the 4 dilation matrices, operating on the original image in Fig. 1.

Definition 1. *Downsampling with matrix D :*

$$y = x \downarrow D \iff y(k) = x(Dk)$$

When we downsample an image with D , we eliminate the pixels having their index in coset Γ_1 . We keep the pixels $x(j) = x(j_1, j_2)$ standing on coset Γ_0 , that is, $j = Dk$, and relocate them at position $k = D^{-1}j$.

In the case of the first 3 dilation matrices, we have eliminated the pixels on the black squares of the chessboard. In the case of D_1 , the image is contracted and reflected. It has to be downsampled twice to recover its original position. In the case of D_2 , the image is contracted and rotated. In the case of D_3 , the image is contracted in one direction, and elongated in the other. In the case of D_4 , the image is contracted in the direction of one of the axes, and reflected on line $x_2 = x_1$.



Fig. 1. Original image

In the first 2 cases, the image is reflected or rotated, but the operation is an isotropy. Therefore, the distance between any 2 points in the image is maintained. In the case of D_3 , the image is severely distorted. In the case of D_4 , it has been contracted along one of the axes; in the orthogonal direction it maintains its original size.

Definition 2. *Upsampling with matrix D :*

$$y = x \uparrow D \iff y_k = \begin{cases} x_r & \text{if } k = Dr \\ 0 & \text{else} \end{cases} .$$

When an image is downsampled, and afterwards upsampled, it recovers its initial size and orientation. The pixels that were eliminated in the downsampling process are zero, and marked in black in Fig. 4, where a detail of the images is shown. The influence of the different cosets can be clearly appreciated.

6 The Wavelet Transform

Let $c^{(0)}$ be the original image. Let $f_0(x)$ be the function associated to the original image, $f_0 \in V_0$. If we decompose $f_0(x)$ into the sum of its projections on V_{-1} and W_{-1} , we have

$$\begin{aligned} f_0(x) &= \sum_{k \in \mathbb{Z}^2} c_k^{(0)} \Phi(x - k) \\ &= f_{-1}(x) + r_{-1}(x) \\ &= \sum_{k \in \mathbb{Z}^2} c_k^{(-1)} \frac{1}{\sqrt{|D|}} \Phi(D^{-1}x - k) + \sum_{k \in \mathbb{Z}^2} d_k^{(-1)} \frac{1}{\sqrt{|D|}} \Psi(D^{-1}x - k). \end{aligned}$$

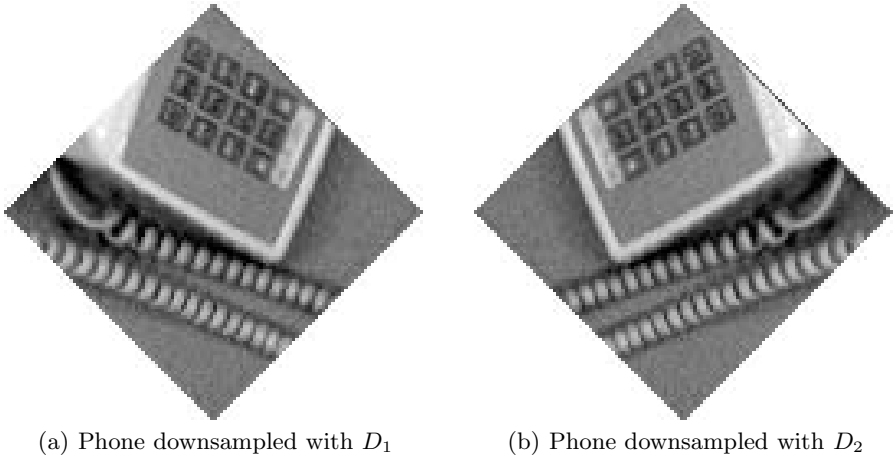


Fig. 2. Effects of downsampling

Applying the orthogonality condition, and equations 1 y 2, we deduce the formulae for analysis

$$c_k^{(-1)} = \frac{1}{\sqrt{|D|}} \sum_{j \in Z^2} h_{(j-Dk)} c_j^{(0)}, \quad (5)$$

$$d_k^{(-1)} = \frac{1}{\sqrt{|D|}} \sum_{j \in Z^2} g_{(j-Dk)} c_j^{(0)}, \quad (6)$$

which enable us to calculate the coefficients of one step of the wavelet transform. Applying the same formulae to $c_k^{(-1)}$ instead of $c_k^{(0)}$, we have 2 steps of the transform.

These steps are shown in Figure 5, where we have applied the Kovacevic-Vetterli [4] which operates with D_1 . The detail coefficients are very small and had to be rescaled to appreciate the details.

In a similar way, we obtain the synthesis formula

$$c_k^{(0)} = \frac{1}{\sqrt{|D|}} \left[\sum_{j \in Z^2} h_{(k-Dj)} c_j^{(-1)} + \sum_{j \in Z^2} g_{(k-Dj)} d_j^{(-1)} \right] \quad (7)$$

Formulae 5 and 6 can be understood as convolutions of the image with filters h and g , (where g indicates the reverse of filter g), followed by downsampling with D , whereas formula 7 indicates upsampling with D , followed by convolutions with h or g , summed up.

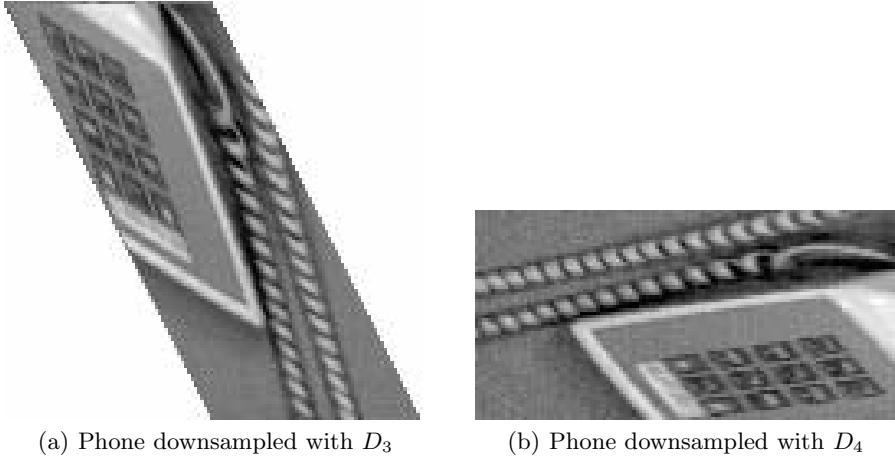


Fig. 3. Effects of downsampling

7 Condition on Eigenvalues Versus Singular Values

It is not desirable to have distortions in an image. A matrix is an isotropy if its singular values are equal; that is why D_1 and D_2 work well.

In addition, a dilation matrix must be an expansion of the plane. We claim that it not enough to have an expansion in any norm, we need to have an expansion in Euclid norm, that corresponds visually to our notion of distance.

Lemma 1. *If the singular values of D verify $\sigma_k > 1$ for $k = 1, 2$, then D is an expansion of the plane in Euclid norm, and D^{-1} is a contraction in the same norm.*

Proof. By the singular value decomposition of matrix D , there exist two 2×2 orthogonal matrices U and F such that

$$U^T D F = S = \begin{bmatrix} \sigma_1 & 0 \\ 0 & \sigma_2 \end{bmatrix},$$

where

$$\sigma_k = \sqrt{\lambda(D^T D)},$$

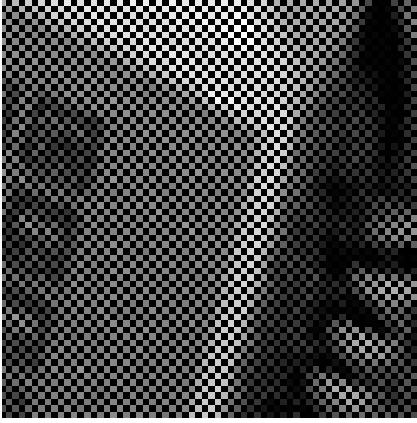
and

$$\sigma_1 \geq \sigma_2 > 1$$

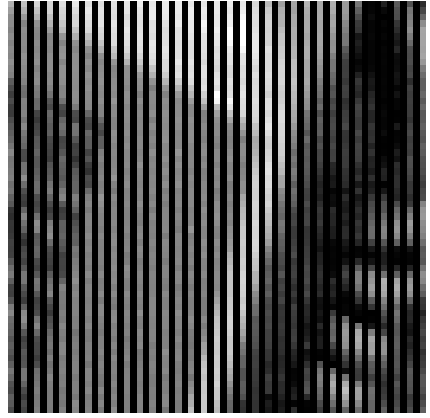
by hypothesis. Then

$$D = U S F^T,$$

$$D^{-1} = F S^{-1} U^T,$$



(a) Phone downsampled and upsampled with D_1 , D_2 , or D_3 .



(b) Phone downsampled and upsampled with D_4

Fig. 4. Effect of subsampling followed by upsampling

so that

$$\|D^{-1}\|_2 = \max_k \frac{1}{\sigma_k} = \frac{1}{\sigma_2} < 1.$$

Then

$$\begin{aligned} \|x\|_2 &= \|D^{-1}Dx\|_2 \leq \|D^{-1}\|_2 \|Dx\|_2, \\ \sigma_2 \|x\|_2 &= \|D^{-1}\|_2^{-1} \|x\|_2 \leq \|Dx\|_2 \end{aligned}$$

and D is an expansion of the plane in Euclid norm. Besides,

$$\|D^{-1}x\|_2 \leq \|D^{-1}\|_2 \|x\|_2 = \frac{1}{\sigma_2} \|x\|_2,$$

and D^{-1} is a contraction in Euclid norm.

Lemma 2. Let λ_k be the eigenvalues of a real $n \times n$ matrix A , and σ_k its singular values. If $\sigma_k > 1$, $\forall k$, then $|\lambda_k| > 1$, $\forall k$.

Proof. Since $A^T A$ is a symmetric positive definite matrix, the Rayleigh quotient verifies:

$$\frac{x^T A^T A x}{x^T x} \geq \sigma_m^2 \quad \forall x \quad (8)$$

where σ_m^2 is the smallest eigenvalue of $A^T A$, and by hypothesis $\sigma_m^2 > 1$. Let x be the eigenvector of A corresponding to λ_k ; replacing in Eq. 8), we obtain $\lambda_k^2 > 1$. Then $|\lambda_k| > 1 \quad \forall k$.

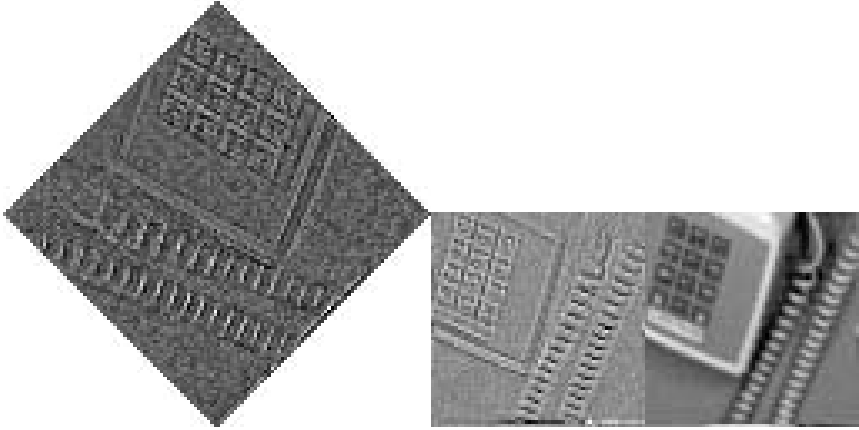


Fig. 5. 2 steps of the KV transform: $d_k^{(-1)}$, $d_k^{(-2)}$ and $c_k^{(-2)}$

Remark 2. The inverse is not true. For example, matrix

$$A_1 = \begin{bmatrix} 0.6 & 0.3 \\ 3.0 & -0.6 \end{bmatrix}$$

has eigenvalues $\lambda = \pm 1.1225$, and singular values $\sigma_1 = 3.1057$, $\sigma_2 = 0.4057$.

Thus we have shown that it is not enough that the eigenvalues be larger than 1 in absolute value to guarantee that the matrix is an expansion in 2 norm.

8 The Elementary Set U

Let $L = \{e_0, e_1\} = \{(0, 0), (1, 0)\}$ be a set of representatives of group Z^2/DZ^2 .

Let f be the function

$$f : \mathfrak{R}^2 \rightarrow \mathfrak{R}$$

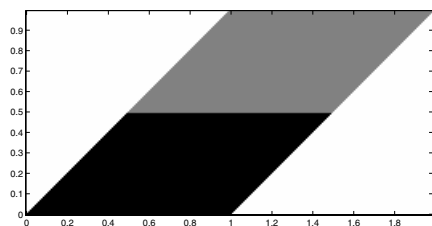
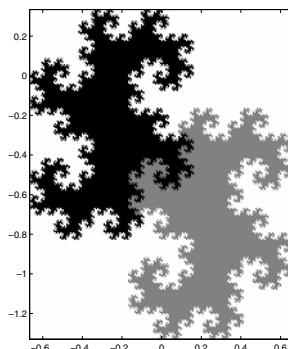
$$f(x) = f(Dx - e_0) + f(Dx - e_1).$$

It can be proved that there exists a compact set $U = U(D, L)$ in \mathfrak{R}^2 — see [2], of which f is the characteristic function, that satisfies

$$U = \{D^{-1}(U + e_0)\} \cup \{D^{-1}(U + e_1)\} = D^{-1}(U + L), \tag{9}$$

meaning that the elementary set is the disjoint union of 2 reduced copies of itself. In Figs. 6 and 7 we have drawn the elementary sets for D_1 , D_2 and D_4 . We have colored subset $D^{-1}U$ in black, and subset $D^{-1}(U + e_1)$ in grey. The elementary set for D_3 is an interval: $[0 \ 1]$.

In the case of D_1 , the elementary set is a parallelogram; in the of D_2 , the elementary set is the so-called twin dragon, and in the case of D_4 , it is a square. To determine the continuity of the wavelet, a tessellation of the plane with these sets has to be done. With matrix D_3 no continuous wavelet can be constructed on the plane.

(a) Elementary set U for D_1 .(b) Elementary set U for D_2 .**Fig. 6.** Elementary sets associated to D_1 and D_2 **Fig. 7.** Elementary set associated to $D = D_4$

9 Conclusions

We have introduced the nonseparable wavelet transform, and show its visual effects. Most of these effects are a consequence of the choice of the dilation matrix. We have analyzed several dilation matrices, and shown how some of them produce distortions.

We have explained why requirements on the singular values of a dilation matrix are better than the requirements on its eigenvalues; simple conditions give an expansion in Euclid norm, and ensure that they will not produce distortions. We have also drawn the elementary set determined by the matrices analyzed.

References

- [1] Skodras, A., Christopoulos, C., Ebrahimi, T.: Jpeg2000: The upcoming still image compression standard. Elsevier, Pattern Recognition Letters **22** (2001) 1337–1345
- [2] Cohen, A., Daubechies, I.: Non-separable bidimensional wavelet bases. Revista Matematica Iberoamericana **9** (1993) 51–137
- [3] Karoui, A., Vaillancourt, R.: Nonseparable biorthogonal wavelet bases of $L^2(\mathbb{R}^n)$. CRM Proceedings and Lecture Notes American Math. Society **18** (1999) 135–151

- [4] Kovacevic, J., Vetterli, M.: Nonseparable multidimensional perfect reconstruction filter banks and wavelet bases for R^n . *IEEE Trans. Inf. Theor.* **38** (1992) 533–555
- [5] Cabrelli, C., Heil, C., Molter, U.: Accuracy of lattice translates of several multidimensional refinable functions. *J. of Approximation Theory* **95** (1998) 5–52
- [6] Cabrelli, C., Heil, C., Molter, U.: Polynomial Reproduction by Refinable Functions. Ka-Sing Lau (1999)
- [7] Ruedin, A.M.C.: Construction of nonseparable multiwavelets for nonlinear image compression. *Eurasip J. of Applied Signal Proc.* **2002**, issue **1** (2002) 73–79
- [8] Heil, C., Colella, D.: Dilation Equations and the Smoothness of Compactly Supported Wavelets. J. Benedetto and M. Frazier, editors, CRC Press (1994)
- [9] Ayache, A.: Construction of non-separable dyadic compactly supported orthonormal wavelet bases $L^2(R^2)$ of arbitrarily high regularity. *Revista Mat. Iberoamericana* **15** (1999) 37–58
- [10] Kovacevic, J., Vetterli, M.: New results on multidimensional filter banks and wavelets. *Proc. IEEE Int. Symposium on Circuits and Systems* (1993)
- [11] He, W., Lai, W.: Examples of bivariate non-separable continuous compactly supported orthonormal wavelets. *IEEE Trans. on Image Processing* **9** (2000) 949–953
- [12] Faugère, J.C., de Saint-Martin, F.M., Rouillier, F.: Design of regular nonseparable bidimensional wavelets using grobner basis techniques. *IEEE Trans. on Signal Processing* **46** (1998) 845–856
- [13] Belogay, E., Wang, Y.: Arbitrarily smooth orthogonal nonseparable wavelets in R^2 . *SIAM J. Math. Anal.* **30** (1999) 678–697
- [14] Ruedin, A.: Nonseparable orthogonal multiwavelets with 2 and 3 vanishing moments on the quincunx grid. *Proc. SPIE Wavelet Appl. Signal Image Proc. VII* **3813** (1999) 455–466
- [15] Ruedin, A.M.C.: Balanced nonseparable orthogonal multiwavelets with two and three vanishing moments on the quincunx grid. *Wavelet Appl. Signal Image Proc. VIII, Proc. SPIE* **4119** (2000) 519–527
- [16] Entezari, A., Moller, T., Vaisey, J.: Subsampling matrices for wavelet decompositions on body centered cubic lattices. *IEEE Sign. Proc. Lett.* **11** (2004) 733–735

Evolutionary Tree-Structured Filter for Impulse Noise Removal

Nemanja I. Petrović¹ and Vladimir S. Crnojević²

¹ Ghent University, Dept. of Telecommunications and Information Processing
(TELIN), IPI, Sint-Pietersnieuwstraat 41, B-9000 Gent, Belgium

petra@uns.ns.ac.yu

² Faculty of Engineering, Trg Dositeja Obradovića 6,
21000 Novi Sad, Serbia

crnojevic@uns.ns.ac.yu

Abstract. A new evolutionary approach for construction of uniform impulse noise filter is presented. Genetic programming is used for combining the basic image transformations and filters into tree structure, which can accurately estimate noise map. Proposed detector is employed for building switching-scheme filter, where recursively implemented α -trimmed mean is used as the estimator of corrupted pixel values. The proposed evolutionary filtering structure shows very good results in removal of uniform impulse noise, for wide range of noise probabilities and different test images.

1 Introduction

One of the common problems in image processing is impulse noise. It is often generated as a result of acquisition or transmission errors [1]. Relatively small amount of this noise can severely degrade image quality. Also, it can be a cause of serious problems in further processing, especially when linear techniques are applied. Hence, impulse noise filtering is a required preprocessing stage and a number of nonlinear techniques have been developed in order to treat this problem properly.

Regardless of the origin of image impulse noise, it has two major properties: only certain percentage of image pixels is contaminated, and the intensity of a damaged pixel is significantly different when compared to intensity of the pixels in its neighborhood. Therefore image containing impulse noise \mathbf{x} can be described as follows:

$$x_{ij} = \begin{cases} n_{ij}, & \text{with probability } p \\ z_{ij}, & \text{with probability } 1 - p \end{cases}, \quad (1)$$

where z_{ij} denotes noiseless image pixel at the location (i, j) , $n_{ij} \in [N_{\min}, N_{\max}]$ are noisy impulses and p is noise probability. Frequently used impulse noise model is salt-and-pepper noise, where noisy pixels take either N_{\min} or N_{\max} value. In this paper we focus on more general random-valued or uniform impulse noise model, where n_{ij} can take any value within the given interval $([0, 255])$ for

grayscale images), with uniform probability distribution. Differences between values of noisy pixel and noise-free pixels in its local neighborhood are more significant for salt-and-pepper noise than for uniform impulse noise. As a result, removal of uniform impulse noise is more challenging task and the majority of existing filters fail with this noise model.

Nonlinear techniques, used for impulse noise removal, are mainly based on median and its modifications [2]. These robust estimators are resistant to high concentrations of impulse noise. On the other hand, they are prone to destroying fine image details. If the filter is applied uniformly across the image, this blurring effect can be very strong and noise-free pixels are unnecessarily modified. Consequently, modern impulse noise filters use switching scheme mechanism [3], consisted of detector, whose task is to identify noisy impulses, and estimator which can be any of classical robust estimators. Only pixels detected as noisy will be replaced by estimated values. Performances of such filters are mainly influenced by quality of the detector. Good detector has to meet two opposing demands: to suppress an impulse noise and to preserve fine details. Contemporary algorithms [2]-[7] try to solve this problem by analyzing difference between value of processing pixel and one or more local neighborhood statistics. Differences are compared to corresponding thresholds and decision if pixel is noisy or noise-free is based on the results of these comparisons.

Recently, an impulse detection method based on pixel-wise MAD (PWMAD) is proposed for impulse noise removal [8]. It identifies noisy pixels by iterative removal of image details. Result of these consecutive image details filtering is map of image regions with no details, but only significant variations in illumination related to noise. Transition from this map to noise map is straightforward.

In this work, the approach to detector design used in PWMAD is extended. Instead of ad-hoc construction where several iterations of median filtering are used, genetic programming (GP) [9], relatively recent and fast developing approach to automatic programming, is employed to construct optimal filter from a broader set of simple filters. In GP, solution to a problem is represented as a computer program in the form of a parse tree, consisting of primitive functions and terminals. It has been chosen among other learning algorithms due to its capability to fit extremely nonlinear functions easily and mitigate influence of futile features. GP employs evolutionary principles of natural selection and recombination to search the space of all possible solutions (combinations of filters) in order to find the most satisfactory one. The obtained GP tree should transform noisy image into noise map. GP has been chosen among other learning algorithms for several reasons:

- The obtained tree is composed of simple building blocks – common image processing filters and operations,
- GP evolution is capable of rejecting irrelevant primitive filters, and keeping only those that contribute significantly to overall filtering performance,
- The structure of the GP tree can be analyzed in order to bring conclusions about the influence of its building blocks,
- GP has capability to fit extremely nonlinear functions easily.

Tree structures, which are used for transforming noisy image into noise map, are trained in such a manner that iterative filtering effectively suppresses noise while preserves image details from blurring. Attention is also given to estimator construction. Instead of using simple median, recursive α -trimmed mean is employed for improvement of the filtering performance.

This paper is organized as follows. The design method of an impulse detector is presented in Section 2, while the practical solution found by this new approach is presented in Section 3. Section 4 contains analysis and comparisons of the new filter performances and finally, conclusions are drawn in Section 5.

2 Proposed Method

Proposed approach is based on standard switching scheme design. Let x_{ij} and y_{ij} denote pixels with coordinates (i, j) in a noisy and a filtered image, respectively. If the estimated value of a particular noisy image pixel is $\varphi(x_{ij})$, then the filtered image is defined as

$$y_{ij} = \varphi(x_{ij})M_{ij} + x_{ij}(1 - M_{ij}), \quad (2)$$

where M_{ij} is the binary noise map, containing ones at the positions detected as noisy and zeros elsewhere.

2.1 Detector Design

Detection process is divided into two steps. Firstly, noisy image is transformed through the filtering tree-structure. Afterwards, result of this transformation is compared to threshold T_d . Values larger than threshold are considered as detected noise impulses and designated as “1” in noise map. Values smaller than threshold, denoted as “0” in noise map, indicate noise-free locations.

It has already been shown that it is possible to construct complex filtering structures automatically by means of genetic programming [12]. In this approach GP tree is built from simple image filters and operations presenting GP primitive functions. Tree-structure for noise detection is constructed in a similar way, where common impulse noise filters are used as primitive functions (e.g. median, centre-weighted median, Min and Max, etc.). For purpose of transforming noisy image into noise map, the terminal set consists of only one element – input noisy image, while the primitive function set is composed of standard one-input filters and simple two-input operations.

Terminal and Function Set. Let W^K denote rectangular window centered at the position (i, j) , where the size of the window is $(2h + 1) \times (2h + 1)$ and $K = 2h + 1$. A set of pixels contained in the window W^K , centered at the position (i, j) , is defined as:

$$W_{ij}^K = \{x_{ij} | -h \leq i \leq h, -h \leq j \leq h, K = 2h + 1\}. \quad (3)$$

Each one-input image filter transforms input image x_{ij} into output image y_{ij} . If the filter is two-input then x_{ij} is replaced by x_{1ij} and x_{2ij} . Image pixels can

take an integer value in range $[0, V_{\max}]$, where $V_{\max} = 255$. Filter output y_{ij} is set to 0 if it is smaller than zero, and to V_{\max} if it is larger than V_{\max} . Definition of filters used as primitive functions in genetic programming, along with their GP coding notation, are given as follows:

1. Absolute deviations from the median $\{\mathbf{d_med3x3}, \mathbf{d_med5x5}\}$

$$y_{ij} = |x_{ij} - \mathit{median}(W_{ij}^3)|, \quad y_{ij} = |x_{ij} - \mathit{median}(W_{ij}^5)|, \quad (4)$$

2. Absolute deviations from centre-weighted median $\{\mathbf{d_cwm3x3}, \mathbf{d_cwm5x5}\}$

$$y_{ij} = |x_{ij} - \mathit{median}(W_{ij}^3) \diamond 2k|, \quad k = 2, \quad (5)$$

$$y_{ij} = |x_{ij} - \mathit{median}(W_{ij}^5) \diamond 2k|, \quad k = 6,$$

where $2k$ weight is given to central pixel and symbol \diamond represents repetition operation,

3. Minimum and Maximum $\{\mathbf{Min}, \mathbf{Max}\}$

$$y_{ij} = \min\{W_{ij}^3\}, \quad y_{ij} = \max\{W_{ij}^3\}, \quad (6)$$

4. Inversion $\{\mathbf{Inv}\}$

$$y_{ij} = V_{\max} - x_{ij}, \quad (7)$$

5. Bounded sum and product $\{\mathbf{BoundedSum}, \mathbf{BoundedProd}\}$

$$y_{ij} = x_{1ij} + x_{2ij}, \quad y_{ij} = x_{1ij} + x_{2ij} - V_{\max}, \quad (8)$$

6. Logical sum and product $\{\mathbf{LogicSum}, \mathbf{LogicProd}\}$

$$y_{ij} = \max\{x_{1ij}, x_{2ij}\}, \quad y_{ij} = \min\{x_{1ij}, x_{2ij}\}, \quad (9)$$

7. Algebraic sum and product $\{\mathbf{AlgebraicSum}, \mathbf{AlgebraicProd}\}$

$$y_{ij} = x_{1ij} + x_{2ij} - \frac{x_{1ij} \cdot x_{2ij}}{V_{\max}}, \quad y_{ij} = \frac{x_{1ij} \cdot x_{2ij}}{V_{\max}}. \quad (10)$$

Note that first three filter types are based on standard median, center weighted median and simple morphological filters. An idea was to use these filters to capture important information for impulse noise detection. The purpose of the other operators was to combine that information with logical and arithmetic functions.

Fitness function. The fitness function is based on similarity measure between ideal and calculated noise map for training image. It is composed of two factors: *sensitivity* and *specificity*. If noisy pixels are defined as positive samples and noise-free pixels as negative samples the fitness is given as

$$\mathit{fitness} = \frac{1}{3} (\mathit{sensitivity} + (2 \times \mathit{specificity})) = \frac{1}{3} \left(\frac{t_pos}{pos} + 2 \frac{t_neg}{neg} \right). \quad (11)$$

The numbers of positive and negative samples are denoted as pos and neg , while the numbers of true positive and true negative samples are denoted as t_pos and t_neg . Fitness definition guaranties that evolution process will lead in the direction of the solutions which are more accurate in detecting noise-free pixels than noisy impulses. Thus, when the image is affected by small percent of noisy impulses, uncorrupted pixels will not be needlessly filtered. Furthermore, for larger impulse concentrations in image, filter can be applied multiple times in order to make better results.

GP parameters. Evolutionary process was performed with the following genetic parameters: crossover 55%, mutation 40%, reproduction 5%, population size 200, number of generations 100. Maximal number of nodes in a tree was set to 40 in order to prevent bloat and avoid overfitting. New generations were created by using tournament selection. This method chooses each parent by randomly drawing a number of individuals from the population and selecting only the best of them. Survival method used was halfelitism, where half of the new population was occupied by the best individuals chosen from both parents and children. The remaining places were occupied by the best children still available.

2.2 Estimator

Filtered image has been generated according to switching scheme defined by (2). Instead of using simple median as the estimator, recursive α -trimmed mean has been utilized. It can be defined as a function of the trimming parameter α

$$X_\alpha = \frac{1}{N - 2[\alpha N]} \sum_{i=[\alpha N]+1}^{N-[\alpha N]} X_{(i)}, \quad (12)$$

where $[\cdot]$ is the greatest integer function and $X_{(i)}$ represents the i -th data item in the sorted sample ($X_{(1)} \leq X_{(2)} \leq \dots \leq X_{(N)}$). When $\alpha = 0$, X_α is the sample mean, and when $\alpha = 0.5$, the previous equation reduces to the sample median. Although median is a robust estimator with breakdown point $\epsilon^* = 0.5$ [10], the reason for using less robust α -trimmed mean is found in a fact that recursive filter implementation make estimator safe from such extreme situation when 50% of samples are outliers. Besides, recursive estimator structure will not cause excessive blurring [11] due to existence of noise detector which adaptively selects filtering operation. Therefore, usage of recursive α -trimmed mean produce better noise attenuation and generally better results.

3 Tree-Structured Detector

Training set was built from standard test image *Couple*. Firstly, image was corrupted by a uniform impulse noise, and the information of noise map was recorded. Output of each automatically generated filter is compared to a constant threshold $T_d = 63$, presenting one quarter of full interval of allowed grayscale

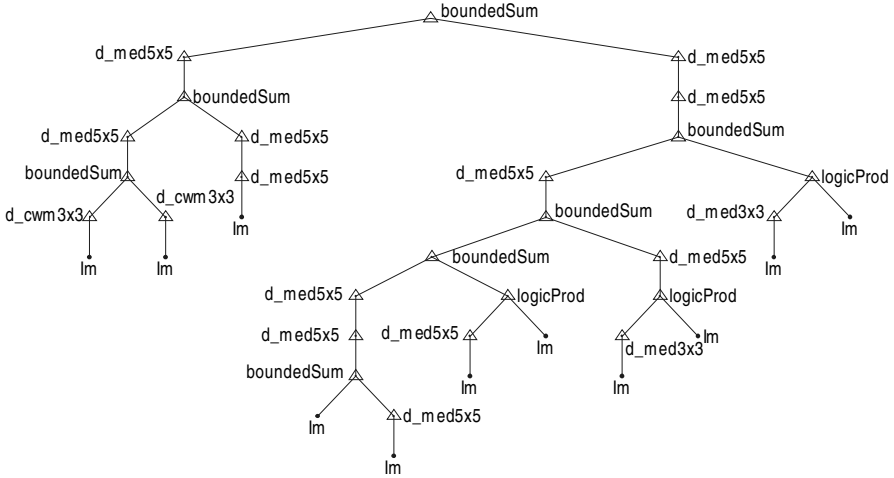


Fig. 1. Tree-structured filter for impulse noise detection

range. Values larger than the threshold are labeled as corrupted in noise map, while smaller values are labeled as noise-free.

For classification problems, it would appear natural to choose half of the interval as a threshold value. Nevertheless, in this particular application, a quarter of the interval is chosen due to the fact that the output will depend on certain combination of absolute differences between actual pixel intensity and some local neighborhood statistics, which, most likely, will not be very large in amplitude for uniform impulse noise case.

An example of tree-structured filter, which is found during the evolutionary process is given in Fig.1, where *Im* denotes input noisy image, and filters are denoted as above. The corresponding LISP syntax string is given as follows:

```
BoundedSum(d_med5x5(BoundedSum(d_med5x5(BoundedSum
(d_cwm3x3(Im), d_cwm3x3(Im))), d_med5x5
(d_med5x5(Im))), d_med5x5(d_med5x5(BoundedSum
(d_med5x5(BoundedSum(BoundedSum(d_med5x5(d_med5x5
(BoundedSum(Im, d_med5x5(Im))), LogicProd
(d_med5x5(Im), Im)), d_med5x5(LogicProd
(d_med3x3(Im), Im))), LogicProd(d_med3x3(Im), Im))))))
```

4 Results

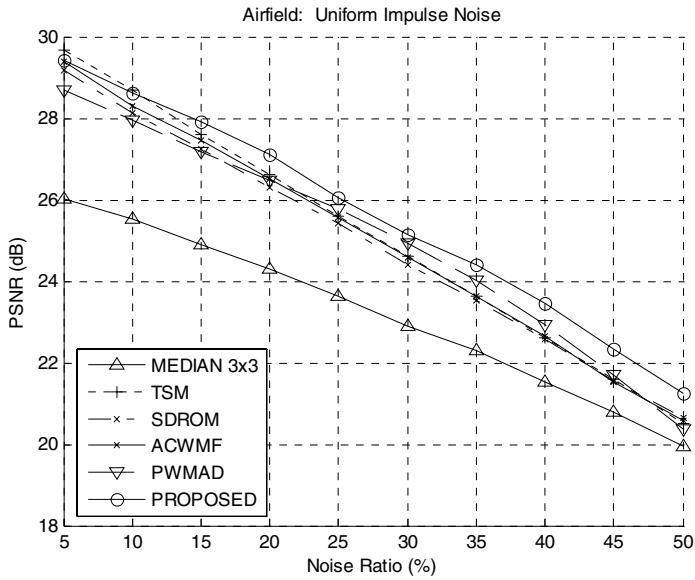
The proposed filter has been compared to some of the standard impulse noise filters. Simulations were made on several standard grayscale test images (reso-

Table 1. Comparative results of impulse noise filters in PSNR (dB). Test images not included in the training set are corrupted by 20% and 40% of uniform impulse noise.

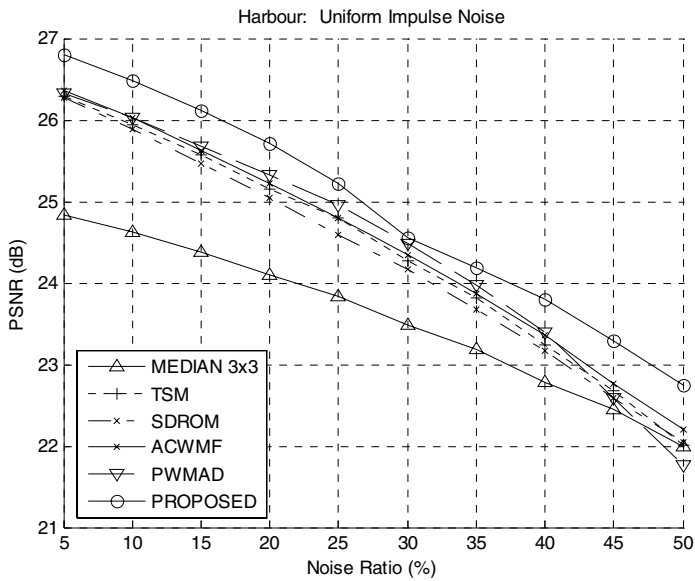
	20%						
Filter	Lena	Bridge	Goldhill	Peppers	Airplane	Boats	Mandrill
Median 3×3	30.82	24.57	29.23	30.85	29.71	28.59	22.04
TSM [4]	33.53	26.89	31.92	33.58	32.13	31.01	23.63
PWMAD [8]	34.38	26.95	32.10	34.43	32.54	31.34	24.32
SDROM [5]	33.76	27.18	32.13	33.61	32.45	31.28	23.83
ACWMF [7]	33.98	26.91	32.12	33.91	32.63	31.25	23.85
Proposed	34.37	27.29	32.30	34.45	32.77	31.56	24.45
	40%						
Filter	Lena	Bridge	Goldhill	Peppers	Airplane	Boats	Mandrill
Median 3×3	26.73	22.31	26.55	26.39	25.69	25.37	21.02
TSM [4]	27.55	23.38	26.98	27.05	25.82	26.03	21.80
PWMAD [8]	29.06	23.44	27.73	28.47	26.98	26.92	21.79
SDROM [5]	28.07	23.68	27.45	27.66	26.70	26.55	21.81
ACWMF [7]	28.42	23.53	27.66	27.88	26.83	26.67	21.86
Proposed	30.18	24.21	28.87	29.99	27.96	27.90	22.13

lution 512×512) that were not included in the training set, contaminated with uniform impulse noise for wide range of impulse concentrations. Simulations were repeated a number of times, for each particular image and noise probability and averaged results are given in Table1 and Fig.2. Quality measure used for evaluation was the peak SNR. Proposed filter is compared to recursive implementations of standard median (filtering window 3×3), TSM [4], SDROM [5] and ACWMF [7], and iterative implementation of PWMAD [8] filter. Proposed filter is implemented in only one iteration for noise ratios $\leq 25\%$, and in two iterations for noise concentrations $>25\%$. Recursive implementation of algorithm means that the estimate of the current pixel is dependent on the new values of previously processed pixels instead of the old ones, while the iterative implementation means that the algorithm is successively applied on the entire noisy image and on the images which are the outputs of the previous filtering.

Performance comparisons of those filters, over noise ratios from 5% to 50%, for test images *Airfield* and *Harbour*, which were not included in the training set are given in Fig.2. Proposed filter shows similar performances for other test images. Table1 shows results of PSNR comparison for images degraded by uniform impulse noise, where 20% and 40% of the pixels are contaminated in each image. These results have similar tendency for all images, where the proposed filter performs best in most situations. Furthermore, performance improvement which is made by proposed algorithm for higher noise ratios is apparently significant. As an example, filtering results obtained for test image *Lena* corrupted with 40% of uniform impulse noise are shown in Fig.3. It can be noticed that the proposed filter achieves the highest level of impulse noise suppression, while image details are still well preserved.



(a)



(b)

Fig. 2. Performance comparison of filtering algorithms: Test images Airfield and Harbour corrupted by uniform impulse noise from 5% to 50%

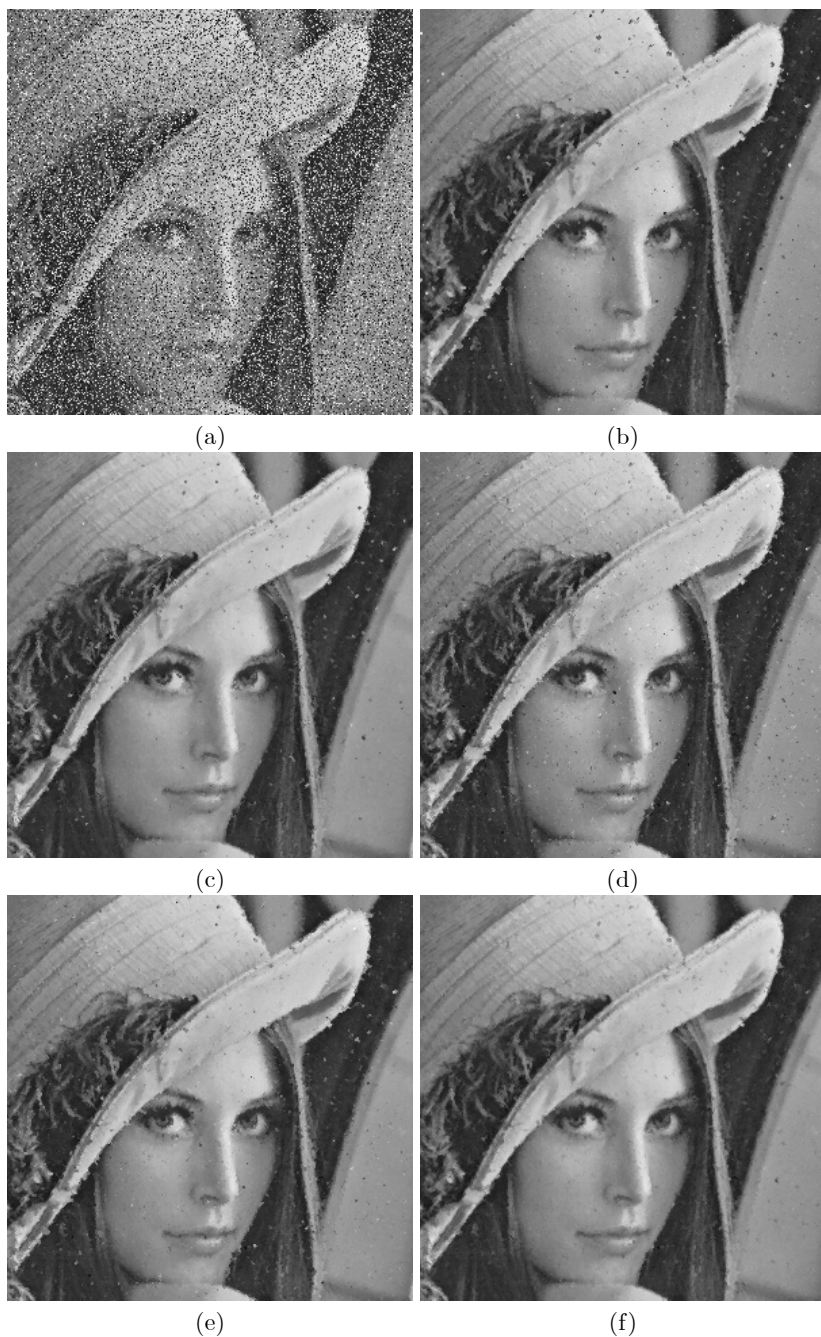


Fig. 3. Enlarged detail of a test image Lena corrupted by 40% of uniform impulse noise: (a) noisy image, (b) TSM, (c) PWMAD, (d) SDROM, (e) ACWMF, (f) proposed

5 Conclusions

A new evolutionary approach to impulse noise filter design is presented. Trade-off between noise suppression and detail preservation is accomplished by using tree-structured impulse detector, trained to accurately detect noise-free pixels. This property of the detector results in possibility to iteratively use the filter for impulse noise removal, without damaging fine details. Results of such procedure are especially noticeable in images corrupted by high noise ratios. Although presented approach requires time consuming training of the impulse detector, obtained tree-structure contains just basic image filters and operations, and can be implemented easily on any platform. Simulations prove that achieved generalization is outstanding. In addition, in most cases the proposed filter outperforms other filters included in comparison. Although large number of filters were included in primitive functions set, GP evolution has chosen only few of them, mostly based on median filter concept. Additional extension of the GP terminals set with randomly generated constants could lead to further improvement of filter performance, by allowing fine adjustment of the threshold T_d . Also, by introducing salt and pepper impulse noise in the training set, the proposed filter can easily be adapted for filtering of fixed-valued impulse noise, which is less demanding task than the uniform impulse noise. Evolutionary approach allowed for the development of high performance filtering structure, that would otherwise be impossible to construct.

References

1. Gonzales, C., Woods, E.: "Digital Image Processing", Prentice-Hall, 2002.
2. Ko, S.-J., Lee, Y.-H.: "Center weighted median filters and their applications to image enhancement," IEEE Trans. Circuits Syst., vol. 38, pp. 984–993, Sept. 1991.
3. Sun, T., Neuvo, Y. : "Detail-preserving median based filters in image processing," Pattern Recognit. Lett., vol. 15, pp. 341–347, Apr. 1994.
4. Chen, T., Ma, K.-K., Chen, L.-H.: "Tri-state median filter for image denoising," IEEE Trans. Image Processing, vol. 8, pp. 1834–1838, Dec. 1999.
5. Abreu, E., Lightstone, M., Mitra, S. K., Arakawa, K.: "A new efficient approach for the removal of impulse noise from highly corrupted images," IEEE Trans. Image Processing, vol. 5, pp. 1012–1025, June 1996.
6. Pok, G., Liu, J.-C., Nair, A. S.: "Selective removal of impulse noise based on homogeneity level information," IEEE Trans. Image Processing, vol. 12, pp. 85–92, Jan. 2003.
7. Chen, T., Wu, H. R.: "Adaptive impulse detection using center-weighted median filters," IEEE Signal Processing Lett., vol. 8, pp. 1–3, Jan. 2001.
8. Crnojevic, V., Senk, V., Trpovski, Z.: "Advanced Impulse Detection Based on Pixel-Wise MAD", IEEE Signal processing letters, vol.11, no.7, July 2004.
9. Koza, J. R. : "Genetic Programming: On the Programming of Computers by Means of Natural Selection", Cambridge, MA: MIT Press, 1992.
10. Huber P.: "Robust Statistics", New York: Wiley, 1981.

11. Nodels, T. A., Gallagher, Jr. N. C.: "Median filters: Some modifications and their properties," IEEE Trans. Acoust., Speech, Signal Processing, vol. ASSP-30, pp. 739-746, May 1982.
12. Aoki, S., Nagao, T.: "Automatic Construction of Tree-structural Image Transformation using Genetic Programming", Proceedings of the 1999 International Conference on Image Processing (ICIP '99), vol. 1, pp. 529 - 533,1999.

Perceived Image Quality Measurement of State-of-the-Art Noise Reduction Schemes

Ewout Vansteenkiste¹, Dietrich Van der Weken²,
Wilfried Philips¹, and Etienne Kerre²

¹ TELIN Dept., Ghent University, Sint-Pietersnieuwstraat 41, Ghent, Belgium

² Dept. of Applied Mathematics, Ghent University, Krijgslaan 281,
S9, Ghent, Belgium

`ervsteen@telin.ugent.be`

Abstract. In this paper we compare the overall image quality of 7 state-of-the-art denoising schemes, based on human visual perception. A psychovisual experiment was set up in which 37 subjects were asked to score and compare denoised images. A perceptual space is constructed from this experiment through multidimensional scaling (MDS) techniques using the perceived dissimilarity and quality preference between the images and the scaled perceptual attributes blurriness and artefacts.

We found that a two-dimensional perceptual space adequately represents the processed images used in the experiment, and that the perceptual spaces obtained for all scenes are very similar. The interpretation of this space leads to a ranking of the filters in perceived overall image quality. We can show that the impairment vector, whose direction is opposite to that of the quality vector, lies between the attribute vectors for blurriness and artefacts, which on their account form an angle of about 35 degrees meaning they do interact. A follow-up experiment allowed us to determine even further why subjects preferred one filter over the other.

1 Introduction

Denoising has been a hot topic for many years in different image processing and analysis tasks, e.g. in image restoration or as a preprocessing step to image segmentation. Multiple advanced denoising schemes have been presented in recent literature using locally adaptive spatial filters in a multi-resolution representation [8, 9, 11], shape-adaptive transforms [4], block-matching with 3D transforms [1], Steerable Filter Pyramid based [5, 10] or Fuzzy Logic [13] techniques.

All of these try to suppress the noise present while preserving as much image content, structures and detail information as possible. Different well-known measures such as the Root Mean Square Error (RMSE) or Peak Signal-to-Noise Ratio (PSNR) are commonly used to compare how well the different filters perform. Although these are good measures to determine a relative distance, for instance to the original noise-free image (if provided), and accordingly to rank the filters, little do these differences tell us about the overall image quality since they don't incorporate human visual information.

Different alternative measures have been proposed tending to incorporate this kind of knowledge, for example the fuzzy similarity measures described in [14]. Yet for certain purposes, e.g. when image distortions are less obvious to be well captured by any instrumental measure, a better approach is to determine a ranking solely based on human visual perception, through some psycho-visual experiment [6].

Here, we perform an experiment on 7 state-of-the-art denoising schemes, trying to rank the filters in perceived overall image quality and to determine why our subjects prefer one filter over the other.

Multidimensional scaling (MDS) analysis has been used in many research areas that deal with the psychovisual evaluation of stimuli, varying in multiple aspects [3]. The rationale underlying this framework is twofold. First, the concept of “homogeneity of perception” should hold, meaning that different subjects are able to reach one common conclusion, e.g. on overall image quality. Secondly, the concept of overall image quality is rarely unidimensional, meaning that different attributes such as noise, blur or artefacts all can influence the perceived quality. The input for the MDS methods are data of psychovisual experiments, and the MDS methods consist of a series of mathematical algorithms that determine the stimulus positions and/or the attribute directions in the MD space on the basis of this input.

In the next sections we will first present a brief overview of the different denoising schemes, Section 2, then elaborate on our psycho-visual experiment and the data processing using MDS, Section 3, before turning to the results and conclusions in Sections 4 and 5.

2 Denoising Schemes

From recent literature the following denoising schemes were selected based on good overall performances. For technical details we refer to the papers in the references:

- **The GOA filter** [13]: A two-step filter where first a fuzzy derivative for eight different directions is computed which is then used to perform a fuzzy smoothing by weighting the contributions of neighboring pixel values. Both stages are based on fuzzy rules using membership functions.
- **The SA-DCT filter** [4]: The Shape-Adaptive DCT scheme uses an over-complete transform-domain filter in conjunction with the anisotropic LPA-ICI technique, which - for every location in the image - adaptively defines an appropriate shape for the transform’s support.
- **The 3D-DFT filter** [1]: The block-matching and 3D filtering approach exploits the possible correlation among similar blocks within an image by filtering in the 3D-transform domain. The third dimension corresponds to stacking together the blocks which are matched as similar.
- **The ProbShrink filter** [8]: This adaptive spatial filter shrinks the wavelet coefficients in a multi-resolution representation according to the probability of the presence of a signal of interest conditioned on a local spatial activity indicator.

- **The BLS-GSM filter** [5]: This method extends filtering in the steerable pyramid domain based on Gaussian Scale Mixtures [9] by employing a two-level (coarse-to-fine) local adaptation to spatial image features.
- **The Bishrink1, Bishrink2 filters** [11]: This method applies a bivariate shrinkage of the wavelet coefficients using the interscale dependencies and the local spatial variance estimation. Two variants were provided corresponding to different noise estimation levels.
- **The SPERRIL filter** [10]: This is an image restoration method, where the regularization (denoising) part is done in the steerable pyramid domain employing the interscale (parent-child) relationships between the coefficients.



(a) Barbara



(b) Face



(c) Hill

Fig. 1. The test scenes used in our psycho-visual experiment

3 Psycho-visual Experiment

3.1 Experimental Setup

A psycho-visual experiment for the assessment of perceived image quality has been described in detail in [6] for images artificially degraded by noise and blur. Partly based on that, we constructed our own experiment that was slightly bigger and focused on more subtle differences.

Stimuli. Three 512×512 pixels 8-bit scenes (Barbara, Face and Hill) containing different kinds of information ranging from texture over fine details to uniform backgrounds, see Fig. 1, were used in the experiment. These images were degraded by additive zero mean white Gaussian noise with a standard deviation of $\sigma = 15$ and were sent to the authors of the filters mentioned above who were asked to denoise them blindly, i.e. without any information on the noise level. The original image, the noisy one $\sigma = 15$, together with the 8 denoised images were presented on the same calibrated display, under comparable lighting conditions. A cut-out example of the Barbara test scene can be seen in Fig. 2.

Method. There were 3 experimental sessions for each scene. In the first session dissimilarity scores and quality preference scores for all pairs of stimuli were collected in a double-stimulus procedure. All unique couples of the 10 stimuli of each scene were displayed on a LCD monitor, one image on the left and the

other on the right. The subjects were instructed to rate the dissimilarity between two images using an integer score between 0 and 5. The subjects were asked not to base that score on any preference, quality or emotional criteria yet. A score of 5 indicated the greatest dissimilarity and a score of 0 implied no perceived difference.

Next, preference scores were asked for image quality on the same couples, this on an integer scale ranging from -3 to +3. Here -3 corresponded to the greatest preference for the left hand image, +3 to the greatest preference for the right hand image, 0 to no preference in perceived quality.

In a second session perceived blur, artefacts and overall quality of the images were judged using a numerical category single-stimulus scaling procedure. Each of the subjects scored the attributes mentioned for all scenes presented separately and the numerical category scale ranged from 0 to 5. The stronger the perceived attribute, the higher the score.

In a third session a follow-up experiment took place in which, based on the outcome of the MDS, 5 well-chosen triples of images were shown to the subjects who were then asked to retain the 2 best images in overall quality and describe in words why they had retained them.

Subjects. 37 subjects took place in the first two sessions of the experiment. 10 additional subjects from the same group took part in the follow-up experiment. Most subjects were familiar with numerical category scaling and the concepts of image quality, blurriness and artefacts. Through a training phase they were made familiar with numerical category scaling. All subjects took part in a trial session involving 10 stimuli from a fourth scene covering the entire range of distortions to adjust the sensitivity of their scale.

3.2 Multidimensional Scaling of Perceived Image Quality

In order to simultaneously model the results from different experiments with the same stimuli and subjects, multi-dimensional geometrical models can be adopted. In such models, images are represented by points in a multi-dimensional space and all observations are related to geometrical properties of these points, such as distances between points and coordinates of point projections onto selected axes. Below, we will describe the class of multidimensional models used in our experiment. We explain the maximum-likelihood optimization criterium used to estimate the model parameters from the experimental data, but start by shortly describing the 3 different kinds of data present: dissimilarity data, preference data and attribute data.

Dissimilarity Data. The first part of the experiment resulted in a 10×10 lower triangular dissimilarity matrix for each subject and each scene. The entry $D_{k,i,j}$ at position (i, j) in the $k = 1 \dots K$ th matrix is the judged dissimilarity for subject k between stimuli i and j . The goal is to construct a stimulus configuration $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$ in a n -dimensional vector space, such that a linear relationship is pursued between transformed dissimilarity scores $TD_{k,i,j} = T_{dk}(D_{k,i,j})$ and the interstimulus distances



(1) BLS-GSM filter



(2) noisy image



(3) original image



(4) ProbShrink filter



(5) SA-DCT filter



(6) GOA filter



(7) SPERRIL filter



(8) BiShrink1 filter



(9) BiShrink2 filter



(10) 3D-DFT filter

Fig. 2. The test images for Barbara as presented in our psycho-visual experiment

$$TD_{k,i,j} \approx d_k \cdot \|\mathbf{x}_i - \mathbf{x}_j\|_2 \quad (1)$$

where the norm is the Euclidean norm and d_k as well as T_{dk} are to be derived from the data. An important implication of this pursued relationship is that the transformed dissimilarities are assumed to be metric, i.e. have ratio properties, since they are compared to metric distances. This implicitly assumes that there exists a monotonic transformation T_{dk} that maps possible non-metric observed dissimilarities $D_{k,i,j}$ into metric transformed dissimilarities $TD_{k,i,j}$ if needed. In our framework to either choosing no transformation at all (when data are metric), a generalized optimum power-like transformation, a generalized Kruskal transformation or an optimum spline transformation, for details see [7].

According to the principle of homogeneity of perception the stimulus configuration should be shared by all subjects, while T_{dk} and d_k can be subject-dependent.

The assumption that the stimulus configuration models the experimental data can be mathematically reformulated by demanding that the error between the transformed dissimilarities and the model predictions is modeled by a chosen probability density function (PDF). Suppose the average transformed model prediction to stimulus pair (i, j) is

$$\widehat{TD}_{k,i,j} = d_k \cdot \|\mathbf{x}_i - \mathbf{x}_j\|_2 \quad (2)$$

then

$$p(TD_{k,i,j}) = \phi[TD_{k,i,j} - \widehat{TD}_{k,i,j}; \sigma_d(k)], \quad (3)$$

is called the link function for subject k and $\sigma_d(k)$ is a measure for the variability in the transformed responses $TD_{k,i,j}$ of subject k on repeated presentations of the same stimulus combination (i, j) . Here ϕ is chosen to be a zero mean generalized Gaussian PDF. The transformation T_{dk} is assumed monotonic, so that the probability $p(D_{k,i,j})$ of obtaining the original dissimilarity $D_{k,i,j}$ can be derived from the probability $p(TD_{k,i,j})$ as

$$p(D_{k,i,j}) = p(TD_{k,i,j}) \cdot |T'_{dk}(D_{k,i,j})| \quad (4)$$

where T'_{dk} denotes the derivative function of T_{dk} . If all subject responses are independent, then

$$P_d = \prod_{k=1}^K \prod_{(i,j)} p(D_{k,i,j}) \quad (5)$$

is the overall probability, according to the model of finding the specific dissimilarity responses $D_{k,i,j}$ that are observed in the experiment. Finding the model parameters that maximize this probability P_d , or equivalently, minimize the inverse of the log-likelihood function

$$L_d = -\log P_d = -\sum_{k=1}^K \sum_{(i,j)} \log p(D_{k,i,j}) \quad (6)$$

corresponds to adjusting the model parameters such that the experimentally observed responses are the most likely outcomes. One can show that the ML criterium is invariant to linear translations and uniform dilation of the stimulus coordinates \mathbf{x}_i . A priori conditions on the stimulus configuration are therefore needed in order to guarantee a unique stimulus configuration, see [6] for details. On the other hand, no measures are taken to uniquely select the orientation of the stimulus configuration, this implies that stimulus configuration can be equal up to a rotational variance.

Preference Data. In a second step we collected 10×10 lower triangular preference matrices for each subject and each scene. The entry $P_{k,i,j}$ at position (i, j) in the $k = 1 \dots K$ th matrix is the judged preference in perceived quality for subject k for stimuli i and j . The ML criterium for the double-stimulus preference data can be derived in a more or less similar way as in the case of dissimilarity data, and is equal to

$$L_p = - \sum_{k=1}^K \sum_{(i,j)} \log p(P_{k,i,j}) \quad (7)$$

where $p(P_{k,i,j})$ and $p(TP_{k,i,j})$ are determined in similar ways as in the dissimilarity case. Different is here that the transformed preference scores $TP_{k,i,j} = T_{pk}(P_{k,i,j})$ are compared against their expected values

$$\widehat{TP}_{k,i,j} = m_k \cdot ([\mathbf{p}_k, \mathbf{x}_i] - [\mathbf{p}, \mathbf{x}_j]). \quad (8)$$

These predictions for subject k are derived from the stimulus positions \mathbf{x}_i , \mathbf{x}_j and the preference vector \mathbf{p}_k and are equal to the vector product

$$[\mathbf{p}_k, \mathbf{x}_i] - [\mathbf{p}, \mathbf{x}_j] = \langle \mathbf{p}_k, \mathbf{x}_i \rangle - \langle \mathbf{p}_k, \mathbf{x}_j \rangle = \langle \mathbf{p}_k, \mathbf{x}_i - \mathbf{x}_j \rangle \quad (9)$$

$$= \sum_{m=1}^n p_{km} \cdot (x_{im} - x_{jm}) \quad (10)$$

between the difference $\mathbf{x}_i - \mathbf{x}_j$, pointing from stimulus j to stimulus i , and the preference vector \mathbf{p}_k . The regression m_k expresses the linear relationship between the transformed scores and the predictions. A consequence of the ratio property (scale invariance) of the transformed preferences is that only the directions of the preference vectors \mathbf{p}_k are uniquely determined, their amplitudes may be scaled arbitrarily.

Very often also, the available preference data can be divided into groups. Here we will consider all 37 quality preferences as belonging to the same group, in which case a single prediction vector $\mathbf{p}_k = \mathbf{p}$ is estimated for all indices in the group. For an elaborate discussion on the estimation of \mathbf{p} we refer to [7].

Attribute Data. Finally, attribute data for blur, artefacts and quality were gathered resulting in 10×1 arrays for each of the subjects, each of the scenes and each of the attributes. The ML criterium for the attribute data is equal to

$$L_a = - \sum_{k=1}^K \sum_{(i,j)} \log p(A_{k,i,j}) \quad (11)$$

where again $p(A_{k,i,j})$ and $p(TA_{k,i,j})$ are derived in similar ways as described above. Different here is that the transformed attribute scores $TA_{k,i,j} = T_{ak}(A_{k,i,j})$ are compared against their expected values

$$\widehat{TA}_{k,i,j} = c_k + f_k \cdot [\mathbf{a}_k, \mathbf{x}_i]. \quad (12)$$

These predictions for subject k are derived from the stimulus positions \mathbf{x}_i and the attribute vector \mathbf{a}_k . The regression parameters c_k and f_k determine the linear relationship between transformed attribute scores and predictions for subject k . Here again the vector-product compares the transformed attribute scores $TA_{k,i,j}$ with the linear prediction

$$c_k + f_k \cdot [\mathbf{a}_k, \mathbf{x}_i] = c_k + f_k \cdot \langle \mathbf{a}_k, \mathbf{x}_i \rangle = c_k + f_k \cdot \sum_{m=1}^n a_{km} \cdot x_{im}.$$

The average strength of attribute k for stimulus i hence increases with the coordinate of the stimulus projection on a one-dimensional axis with scale factor f_k and the direction of the attribute vector \mathbf{a}_k . As in the case of the preference data the attribute data may be subdivided into groups. This corresponds to estimating a single prediction vector $\mathbf{a}_k = \mathbf{a}$ for all indices k in the group. For an elaborate discussion on the estimation of \mathbf{a} again we refer to [7].

3.3 Maximum-Likelihood Estimation

Combining the former sections, in our global experiment the ML criterium

$$L = \sum_{k=1}^K \sum_{(i,j)} \log p(D_{k,i,j}) + \sum_{k=1}^K \sum_{(i,j)} \log p(P_{k,i,j}) + \sum_{k=1}^K \sum_{(i,j)} \log p(A_{k,i,j}) \quad (13)$$

needs to be minimized as a function of the stimulus positions $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_N$, the regression parameters d_k (for dissimilarity), m_k and \mathbf{p}_k (for preference), c_k , f_k and \mathbf{a}_k (for attribute scaling), and the PDF noise standard deviations $\sigma_d(k)$, $\sigma_p(k)$ and $\sigma_a(k)$. In the case of non-metric data, the monotonic transformations T_{dk} , T_{pk} and T_{ak} must also be optimized.

All required optimizations are performed iteratively. If the stimulus configuration needs optimized, then one iteration step involves three stages. In the first stage, the stimulus positions are optimized, assuming fixed values for the regression parameters, the standard deviations and the monotonic transformations. In a second stage, the regression parameters and standard deviations are optimized for a fixed stimulus configuration and known monotonic transformations on the data. Non-metric MDS adds a third stage to each iteration step in which the monotonic transformations are updated.

The following general result on ML estimation can be used to compare the goodness-of-fit of alternative multidimensional models at any stage. The estimation of the different parameters, such as d_k , m_k , c_k and f_k , leads to a number of degrees of freedom (DOF) in the MDS [7]. Suppose now, that L_1 is the optimized ML criterium value for a model with F_1 DOFs, and the $L_2 (> L_1)$ is the

optimized ML criterium value for a reduced model with $F_2 (< F_1)$ DOFs. It can be shown that the statistic

$$G_{12}^2 = 2 \cdot (L_2 - L_1) \quad (14)$$

satisfies a χ^2 -distribution (*chi-squared*) distribution with $F_{12} = F_1 - F_2$ DOFs, i.e., the probability that G_{12}^2 exceeds the value χ^2 is given by

$$P(G_{12}^2 > \chi^2; F_{12}) = \frac{\Gamma(\frac{F_{12}}{2}, \frac{\chi^2}{2})}{\Gamma(\frac{F_{12}}{2})} \quad (15)$$

with Γ the incomplete gamma function [7]. Suppose that $\chi_\alpha^2(F_{12})$ is the value for which $P(G_{12}^2 > \chi_\alpha^2(F_{12}); F_{12}) = \alpha$, then the observed value $G_{12}^2 > \chi_\alpha^2(F_{12})$ is an indication that both models are not equivalent. Indeed, the probability that such a value occurs in the case both models are equivalent is less than α . In our application $\alpha = 0.05$ is chosen.

4 Results

Fig. 3 shows the 2D-geometrical output configuration as optimized by the MDS framework assuming the 37 subjects constitute a homogeneous group, so that the attribute vectors were shared by all. The stimulus configurations were determined for different transformations on the experimental data. There was statistical evidence that a spline interpolation with two kernel knots on the experimental data performed better than a generalized Kruskal or power-like transformation, so that we will use the latter in the subsequent analysis.

Each point in Fig. 3 corresponds to one of the filters shown in Fig. 2. The 95 % confidence intervals on the positions are also plotted as the little ellipses. All stimulus positions were found statistically significant and similar configurations were obtained for all three scenes, apart from a rotational variance and slight change in position of certain filters which we will discuss later.

The arrows point out the directions through which the different attributes should be measured. ‘‘I’’ stands for the impairment vector and is the opposite direction of perceived image quality, the further along the axis the more quality degrades. ‘‘B’’ stands for the perceived blur, the further along the axis the less blur perceived. ‘‘A’’ stands for the artefact axis, the further along the axis the less artefacts perceived.

The orthogonal projection of all points on these axes gives us a relative ranking of the images. Fig. 4 gives us the projection of the perceived quality for all three scenes. Note that for Barbara the ranking seems to be opposite. This is no error in the experiment yet is the result of the rotational variance in the MDS.

The outcome of the different plots in Fig. 4 is comprised in Table 1 where one can see that the original image (3) always comes out best, followed by the 3D-DFT (10) and SA-DCT (5) filter. The BLS-GSM (1) filter, ProbShrink (4) filter and Bishrink 1 (8) follow yet switch places over the different scenes. The GOA filter (6) and SPERRIL filter (7) and Bishrink2 (9) filters are ranked worst, even

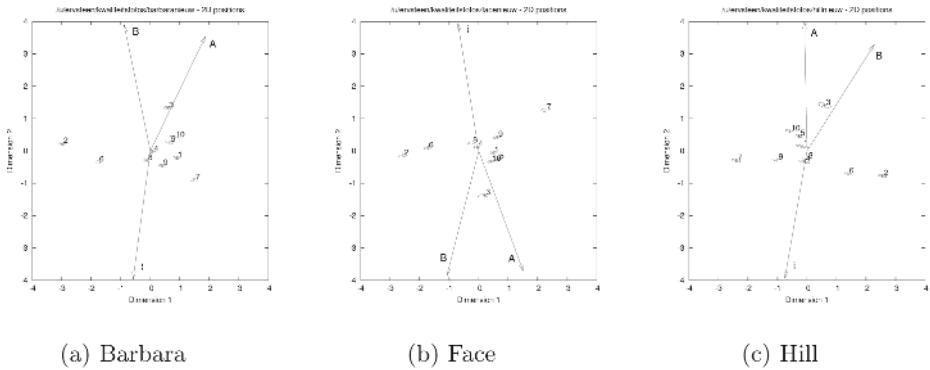


Fig. 3. The Multidimensional scaling output of the experiment. The numbers of the filters correspond to the numbers in Fig. 2.

below the noisy image (2). A connotation to the distances in Fig. 4 can be given through an interpretation of the attribute axes in Fig. 3, as we will show further on. In the same Fig. 3 we can also see that the 3D-DFT (10) and SA-DCT (5) filter, BLS-GSM (1) filter, ProbShrink (4) filter and Bishrink (8) filters seem to cluster, meaning that although they differ in ranking, they do have a common ground in terms of perceived image quality.

As a matter of comparison we also plotted the PSNR-ranking in Table 1. We see that 3D-DFT (10) still performs best in terms of PSNR but now the BLS-GSM (1) filter comes in second in case of Barbara, followed by SA-DCT (5). If we look at the bottom of the table we also see some changes. Next to that, we also notice a bigger shift in ranking through the scenes than in our psychovisual experiment where the top and bottom 3 images were always consistently ranked. Finally, if we e.g. compare the SA-DCT (5) and BLS-GSM (1) filters visually, see cutouts (5) and (1) in Fig. 2, it is clear the SA-DCT filter has less disturbing artefacts (e.g. around the nose), although there is an inverse ranking in PSNR.

Fig. 3 also gives us the direction of the blur and artefacts axis. Projections on those axes can explain the ranking in Table 1. Now since in the middle part of the table the differences are quite small and confidence regions tend to overlap, see Fig. 4, we set up the extra follow-up experiment described in section 3 to try and grasp these smaller differences. The results from this experiment, for the Face image, are presented in Table 2.

This table shows the main attributes taken into account by the different subjects in pointing out the actual difference between the images and should be interpreted as follows: the filter in row i outperforms the filter in column j , mainly based on table entry (i, j) .

For instance, although the 3D-DFT (10) and SA-DCT-filter (5) are very close to one another as well in the 2D-configuration as in the 1D-projection in Fig. 3 and Fig. 4, their main difference lies in the amount of detail information left in 3D-DFT that is not present in SA-DCT. Also we can see that the noisy image is preferred above the Bishrink 2 (9) and GOA (6) filter, mainly because of the blur

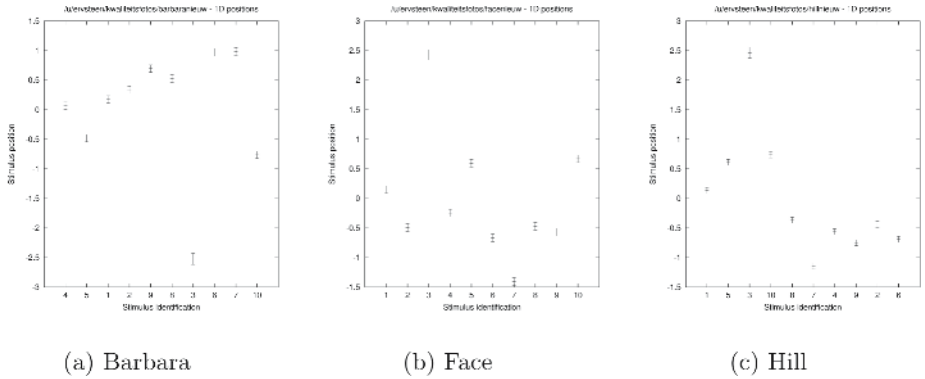


Fig. 4. This figure shows the 1D-geometrical output of the MDS framework for Face. On the X-axis the different filters as numbered in Fig. 2, on the Y-axis the perceived quality.

Table 1. Quality ranking of the filters for each of the 3 scenes based on the outcome of the MDS as well as PSNR. 1 corresponds to best, 10 corresponds to worst.

ranking	Face		Barbara		Hill	
	<i>MDS</i>	<i>PSNR</i>	<i>MDS</i>	<i>PSNR</i>	<i>MDS</i>	<i>PSNR</i>
1 (best)	Original	Original	Original	Original	Original	Original
2	3D-DFT	3D-DFT	3D-DFT	3D-DFT	3D-DFT	3D-DFT
3	SA-DCT	SA-DCT	SA-DCT	BLS-GSM	SA-DCT	SA-DCT
4	BLS-GSM	BiShrink1	Probshrink	SA-DCT	BLS-GSM	BLS-GSM
5	ProbShrink	BLS-GSM	BLS-GSM	Bishrink1	Bishrink1	Bishrink1
6	Bishrink1	Bishrink2	Noisy	ProbShrink	Noisy	ProbShrink
7	Noisy	ProbShrink	Bishrink1	Bishrink2	ProbShrink	Bishrink2
8	Bishrink2	SPERRIL	BiShrink2	SPERRIL	Bishrink2	SRERRIL
9	GOA	Noisy	GOA	Noisy	GOA	Noisy
10 (worst)	SPERRIL	GOA	SPERRIL	GOA	SPERRIL	GOA

in the images. This means that although there is a lot of noise present, subjects tend to prefer the preservation of high frequency information and sharpness of edges in the images. Notice also that not all possible triples were shown, but only those relevant to the quality ranking of Fig. 4.

From this table we can conclude that quality is highly related to the amount of remaining blur in the images. This is also justified by the findings of the MDS where we obtain a direction strongly related to the quality axis.

As for the artefact axis we see from the follow-up experiment it comprises noise and detail information. Since most of the filters perform very well in terms of noise reduction, it is understandable the bigger part will be blurring artefacts and the presence or absence of detailed information.

Nevertheless, from Fig. 3 we consistently see that perceived overall quality is an (inverse) combination of blur and artefacts, which form an angle of about

Table 2. This table shows the main attributes by which the filter in row i is chosen over the filter in column j , based on the follow-up experiment. B = blur, A = artefacts, D = detail information.

	original	3Swdft	SA-DCT	BLS-GSM	Genlik	Noisy	Bishrink2	GOA
original	/	B	B	B	/	N	/	$N + B$
3D-DFT	/	/	D	$B + A$	/	/	/	/
SA-DCT	/	/	/	$A + D$	$B + A$	/	/	/
BLS-GSM	/	/	/	/	$B + D$	/	/	/
Genlik	/	/	/	/	/	B	$B + N$	/
Noisy	/	/	/	/	/	/	B	B
Bishrink2	/	/	/	/	/	/	/	N
GOA	/	/	/	/	/	/	/	/

35 degrees, meaning they are not uncorrelated and the perceived artefacts are mostly blurring artefacts. This is all in accordance with Table 2.

For instance, when we relate the findings in Table 2 to the distances in the 2D-MDS configuration we see that although the noisy image (2) and Bishrink2 (9) filter are very close in perceived overall quality they are quite far apart in the 2D-geometry of Fig. 3. There is a relatively big difference in perceived blur level, see Fig. 3 as well as Table 2, and a smaller difference in perceived artefacts, again see Fig. 3.

The same discussion can be repeated for all 3 scenes, where even the scene content can be taken into account. Notice that although the MDS results are consistent for the bigger part over all of them, some slight changes in ranking might also be explained by the actual image content. This is still work in progress and out of the scope of this paper.

5 Conclusions

The aim of this paper was to compare the perceived image quality of 7 state-of-the-art filters based on a psycho-visual experiment, leading to a ranking that is more true to human visual perception than instrumental measures as the PSNR. We were able to show consistently based on the subjective scores, which filters related best to the original image, namely the 3D-DFT and SA-DCT filters, followed by the BLS-GSM, ProbShrink and Bishrink1 filter. We noticed from the MDS that, although the difference in quality, these five filters seem to cluster more or less, meaning that they show common grounds in terms of human visual perception.

In a follow-up experiment we were able to show why certain filters outperformed others and we could relate this to the findings of the MDS. Blurring and artefacts are shown to be the decision criteria, of which blur seems to carry the biggest load.

As overall conclusions we showed that for these type of filters it is possible to make a consistent ranking based on human visual perception and filters that succeed in denoising images with minor blurring, even though this means leaving

some of the noise present, while introducing minor artefacts, are considered perceptually best. If a trade off needs to be made then the blurring is more a problem than the remaining of noise or artefacts.

References

1. Dabov K., Foi A., Katkovnik V. & Egiazarian K., "Image denoising with block-matching and 3D filtering", *To appear in Image Processing: Algorithms and Systems V, 6064A-30, IST/SPIE Electronic Imaging*, 2006, San Jose, CA., 2006.
2. Donoho D.L. & Johnstone I.M., "Ideal spatial adaptation by wavelet shrinkage", *Biometrika*, vol. 81, no. 3, 1994, pp. 425-455.
3. Escalante-Ramirez B., Martens J.B. & de Ridder H., "Multidimensional Characterization of the Perceptual Quality of Noise-reduced Computed Tomography Images", *J. Visual Comm. Image Representation*, vol. 6, December 1995, pp. 317-334.
4. Foi, A., Dabov K., Katkovnik V., & Egiazarian K., "Shape-Adaptive DCT for Denoising and Image Reconstruction", *Proc. SPIE Electronic Imaging 2006, Image Processing: Algorithms and Systems V, 6064A-18*, San Jose, 2006.
5. Guerrero-Colon J.A. & Portilla J., "Two-Level Adaptive Denoising Using Gaussian Scale Mixtures in Overcomplete Oriented Pyramids", *Proceedings of IEEE ICIP conference*, Genoa, Italy, Sept 2005, pp 105-108.
6. Kayagadde V. & Martens J.B., "Perceptual Characterization of Images Degraded by Blur and Noise: experiments", *Journal of Opt. Soc. Amer. A* 13, June 1996, pp. 1178-1188.
7. Martens J.-B., "Image Technology Design", *Springer*, 2003, Chapter 5.
8. Pizurica A. & Philips W., "Estimating probability of presence of a signal of interest in multiresolution single- and multiband image denoising", "A Joint Inter- and Intrascale Statistical Model for Bayesian Wavelet Based Image Denoising", *IEEE Transactions on Image Processing*, in press.
9. Portilla J., Strela V, Wainwright M & Simoncelli E.P., "Image Denoising using Scale Mixtures of Gaussians in the Wavelet Domain", *IEEE Transactions on Image Processing*, vol. 12, no. 11, 2003, pp. 1338-1351.
10. Rooms F., "Nonlinear Methods in Image Restoration Applied to Confocal Microscopy", *Ph.D thesis*, 2005, Ghent University.
11. Sendur L. & Selesnick I.W., "Bivariate Shrinkage With Local Variance Estimation", *IEEE Signal Processing Letters*, vol. 9, no. 12, 2002, pp. 438-441.
12. Sendur L. & Selesnick I.W., "Bivariate Shrinkage Functions for Wavelet-Based Denoising Exploiting Interscale Dependency", *IEEE Trans. on Signal Processing*, vol 50, no. 11, 2002, pp. 2744-2756.
13. Van De Ville D., Nachtegaal M., Van der Weken D., Kerre E.E., Philips W., Lemahieu I., "Noise Reduction by Fuzzy Image Filtering", *IEEE Transactions on Fuzzy Systems*, vol. 11, no. 4, 2003, pp. 429-436.
14. Van der Weken D., Nachtegaal M. & Kerre E.E., "Using Similarity Measures and Homogeneity for the Comparison of Images", *Image and Vision Computing*, vol. 22 (9), 2004, pp. 695-702.

Multiway Filtering Applied on Hyperspectral Images

N. Renard¹, S. Bourennane¹, and J. Blanc-Talon²

¹ Univ. Paul Cézanne, EGIM, Institut Fresnel (CNRS UMR 6133),
Dom. Univ. de Saint Jérôme, F-13013 Marseille cedex 20, France

² DGA/D4S/MRIS, Arcueil, France

nadine.renard@fresnel.fr, salah.bourennane@fresnel.fr

Abstract. A new multidimensional modeling of data has recently been introduced, which can be used a wide range of signals. This paper presents multiway filtering for denoising hyperspectral images. This approach is based on a tensorial modeling of the desired information. The optimization criterion used in this multiway filtering is the minimization of the mean square error between the estimated signal and the desired signal. This minimization leads to some estimated n -mode filters which can be considered as the extension of the well-known Wiener filter in a particular mode. An ALS algorithm is proposed to determine each n -mode Wiener filter. Using the ALS loop allows to take into account the mode interdependence. This algorithm requires the signal subspace estimation for each mode. In this study, we have extended the well-know Akaike Information Criterion (AIC) and the minimum description length (MDL) criterion to detect the number of dominant eigenvalues associated with the signal subspace. The performance of this new method is tested on hyperspectral images. Comparative studies with classical bidimensional filtering methods show that our algorithm presents good performances.

1 Introduction

Multidimensional model is used in a large range of fields such as data analysis or signal and image processing [1]. In multidimensional signal processing, tensors are built on vector spaces associated with physical quantities such as length, width, height, time, color channels etc... . Each mode of the tensor is associated to a physical quantity. When dealing with data modeled by tensor, the classical processing techniques consist in rearranging or splitting the data set into matrices or vectors in order for the classical algebraic processing methods to be applicable. The original data structure is then rebuilt after processing. In particular, hyperspectral images are split or unfold into observation vectors or matrices in order to apply classical methods based on covariance matrix or cross-spectral matrix. The splitting of multidimensional data reduces considerably the information quantity related to the whole tensor and then the possibility of studying the relations between components of different channels is lost. In this study, hyperspectral images are considered as whole tensor. Hence, we propose a

multiway filtering [2], [3] for denoising hyperspectral images. This new approach implicitly implies the use of multilinear algebra and mathematical tools [4] that extend the SVD to tensors. In [5] a survey on tensor signal filtering for based on a subspace approach is presented.

The paper is organized as follows. Section 2 presents the tensor model and a short overview of its main properties. The tensor formulation of the classical noise-removing problem and new tensor filtering notations are introduced. Section 3 presents a new version of Wiener Filtering based on the n -mode signal subspace and tensor decomposition. Section 4 presents AIC and MDL criteria for estimating the n -mode signal subspace dimension required by tensor filtering. Section 5 details the final ALS algorithm by summarizing step by step the filtering process.

Section 6 contains some comparative results concerning the multiway filtering and channel-by-channel based Wiener filtering, in the case of noise reduction in hyperspectral images. Finally, conclusion is presented in section 7.

2 Tensor of Hyperspectral Images and Multiway Filtering

Hyperspectral images can be modeled by a third order tensor $\mathcal{A} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ (see Figure 1) in which I_1 is the number of rows, I_2 is the number of columns, and I_3 is the number of spectral channels. Each dimension of tensor is called n -mode where n refers to the n^{th} index.

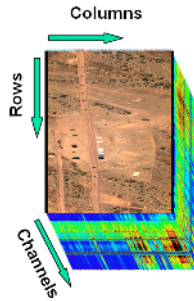


Fig. 1. Tensor of hyperspectral images

In order to study properties of data tensor \mathcal{A} in a given n -mode, let's define $E^{(n)}$ the n -mode vector space of dimension I_n , associated with the n -mode of tensor \mathcal{A} . By definition, $E^{(n)}$ is generated by the column vectors of the n -mode unfolding matrix. The n -mode unfolding matrix A_n of tensor $\mathcal{A} \in \mathbf{R}^{I_1 \times \dots \times I_N}$ is defined as a matrix from $\mathbf{R}^{I_n \times M_n}$, with:

$$M_n = I_1 \cdots I_{n-1} I_{n+1} \cdots I_N. \quad (1)$$

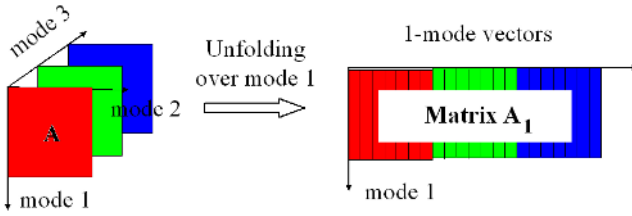


Fig. 2. n -mode unfolding of tensor \mathcal{A}

A_n columns are the I_n -dimensional vectors obtained from \mathcal{A} by varying index i_n and keeping the other indices fixed. These vectors are called tensor \mathcal{A} "n-mode vectors". An illustration of the 1-mode unfolding of a image is represented on Figure 2.

The hyperspectral image \mathcal{X} is assumed as the sum of the desired information with additive white and Gaussian noise \mathcal{B} results in the data tensor:

$$\mathcal{R} = \mathcal{X} + \mathcal{B}. \quad (2)$$

The goal of this study is to estimate the desired signal \mathcal{X} thanks to a multidimensional filtering of the data:

$$\hat{\mathcal{X}} = \mathcal{R} \times_1 H_1 \times_2 H_2 \times_3 H_3,$$

where \times_n is the n -mode product. The n -mode product between a data tensor \mathcal{R} and matrix H_n represents the consecutive matrix products between matrix H_n and the I_n -dimensional vectors obtained from \mathcal{R} by varying index i_n and keeping the other indexes fixed. From a signal processing point of view, the n -mode product is a n -mode filtering of data tensor \mathcal{R} by n -mode filter H_n . Then, H_n is the n -mode filter applied to the n -mode of the data tensor \mathcal{R} , for all $n = 1$ to 3. In the following we establish the expression of the multiway Wiener filtering for tensor of order N [2].

The optimization criterion chosen to determine the optimal n -mode filters $\{H_n, n = 1, \dots, N\}$ is the minimization of the the mean square error between the estimated signal $\hat{\mathcal{X}}$ and the initial signal \mathcal{X} :

$$e(H_1, \dots, H_N) = \mathbb{E} \left(\|\mathcal{X} - \mathcal{R} \times_1 H_1 \cdots \times_N H_N\|^2 \right). \quad (3)$$

In extension of the first order case, n -mode filters H_n correspond to n -mode Wiener filters.

In the classical multidimensional and multi-mode signal processing assumption [6],[7], $E^{(n)}$ is the superposition of two orthogonal subspaces: the signal subspace $E_1^{(n)}$ of dimension K_n , and the noise subspace $E_2^{(n)}$ with dimension $I_n - K_n$, such as $E^{(n)} = E_1^{(n)} \oplus E_2^{(n)}$.

3 Expression of n -Mode Wiener Filters

Developing the squared norm of equation (3), and unfolding it over the n -mode and expressing the tensorial scalar product with the trace operator ($\text{tr}(\cdot)$) lead to [2]:

$$e(H_1, \dots, H_N) = \text{E} \left[\|X_n\|^2 \right] - 2\text{E} \left[\text{tr} \left(g_{XR}^{(n)} H_n^T \right) \right] + \text{E} \left[\text{tr} \left(H_n G_{RR}^{(n)} H_n^T \right) \right], \quad (4)$$

where

$$g_{XR}^{(n)} = X_n \mathbf{q}^{(n)} R_n^T, \quad (5)$$

with

$$\mathbf{q}^{(n)} = H_1 \otimes \dots \otimes H_{n-1} \otimes H_{n+1} \dots \otimes H_N, \quad (6)$$

and

$$G_{RR}^{(n)} = R_n \mathbf{Q}^{(n)} R_n^T, \quad (7)$$

with

$$\mathbf{Q}^{(n)} = \mathbf{q}^{(n)T} \mathbf{q}^{(n)} = H_1^T H_1 \otimes \dots \otimes H_{n-1}^T H_{n-1} \otimes H_{n+1}^T H_{n+1} \dots \otimes H_N^T H_N. \quad (8)$$

The symbol \otimes defines the Kronecker product.

The optimal n -mode Wiener filters $\{H_1, \dots, H_N\}$ are the arguments that minimize the mean square error (4). These filters are found when $\mathbf{grad}(e) = \left[\frac{\partial e}{\partial H_1} \dots \frac{\partial e}{\partial H_N} \right]^T = \mathbf{0}$, that is when $\frac{\partial e}{\partial H_n}$ are simultaneously null for all $n = 1$ to N . Let's study $\frac{\partial e}{\partial H_n}$ for a given n -mode. The n -mode filters H_m are supposed fixed for all $m \in \{1, \dots, N\} - \{n\}$, then $g_{XR}^{(n)}$ and $G_{RR}^{(n)}$ are independent from n -mode filters H_n . Hence imposing $\frac{\partial e}{\partial H_n} = 0$ and extracting H_n lead to the optimal filter which minimizes the mean square error criterion (3) for fixed m -mode filters H_m , $m \neq n$:

$$H_n = \gamma_{XR}^{(n)} \Gamma_{RR}^{(n)-1}, \quad (9)$$

where:

$$\gamma_{XR}^{(n)} = \text{E} \left[g_{XR}^{(n)} \right] = \text{E} \left[X_n \mathbf{q}^{(n)} R_n^T \right], \quad (10)$$

is the $\mathbf{q}^{(n)}$ -weighted covariance matrix between the signal \mathcal{X} and the data \mathcal{R} , and:

$$\Gamma_{RR}^{(n)} = \text{E} \left[G_{RR}^{(n)} \right] = \text{E} \left[R_n \mathbf{Q}^{(n)} R_n^T \right], \quad (11)$$

is the $\mathbf{Q}^{(n)}$ -weighted covariance matrix of the data. In the following, γ refers to the $\mathbf{q}^{(n)}$ -weighted covariance, and Γ to the $\mathbf{Q}^{(n)}$ -weighted covariance.

In order to determine the expression of $\gamma_{XR}^{(n)}$ and $\Gamma_{RR}^{(n)}$ of equation (9), let's make the assumption that the n -mode unfolding matrix of the signal X_n can be expressed as a weighted combination of K_n vectors from the n -mode signal sub-space E_n^1 :

$$X_n = V_s^{(n)} O, \quad (12)$$

with $X_n \in \mathbb{R}^{I_n \times M_n}$, where $M_n = I_1 \cdots I_{n-1} I_{n+1} \cdots I_N$, $V_s^{(n)} \in \text{St}(I_n, K_n)$ the matrix which contains the K_n I_n -dimensional orthogonal vectors of the basis of the n -mode signal sub-space E_n^1 . $O \in \mathbf{R}^{K_n \times M_n}$ is a random weighted matrix whose terms are supposed mutually independent.

Following [2] and after some computations, the final expression of H_n n -mode filter associated to fixed H_m m -mode filters, $m \neq n$, becomes:

$$H_n = V_s^{(n)} \begin{bmatrix} \frac{\lambda_1^\gamma - \sigma_\gamma^{(n)2}}{\lambda_1^\gamma} & & 0 \\ & \ddots & \\ 0 & & \frac{\lambda_{K_n}^\gamma - \sigma_\gamma^{(n)2}}{\lambda_{K_n}^\gamma} \end{bmatrix} V_s^{(n)T}, \quad (13)$$

in which $\{\lambda_i^\gamma, \forall i = 1, \dots, K_n\}$ and $\{\lambda_i^\Gamma, \forall i = 1, \dots, K_n\}$ are the K_n largest eigenvalues, respectively of the matrix $\gamma_{XR}^{(n)}$ and $\Gamma_{RR}^{(n)}$ defined in the relations (10) and (11). Also, $\sigma_\gamma^{(n)2}$ can be estimated by determining the $I_n - K_n$ smallest eigenvalues mean of $\gamma_{RR}^{(n)}$:

$$\sigma_\gamma^{(n)2} = \frac{1}{I_n - K_n} \sum_{i=K_n+1}^{I_n} \lambda_i^\gamma. \quad (14)$$

Note that this expression requires the unknown parameter K_n . To apply it on real data, without *a priori* knowledge, we have to estimate it. We propose in the following section two criteria.

4 n -Mode Signal Subspace Estimation

In order to estimate the signal subspace dimension for each n -mode, we extend the well-know detection criterion [8]. Thus, the optimal signal subspace dimension is obtained merely by minimizing one of AIC or MDL criteria.

The signal subspace dimension is equal to the n -mode rank of the noisy image \mathcal{R} .

Consequently, for each n -mode unfolding of \mathcal{R} , the form of detection criterion AIC can be expressed as

$$AIC(k) = -2N \sum_{i=k+1}^{i=I_n} \log \lambda_i + N(I_n - k) \log \left(\frac{1}{I_n - k} \sum_{i=k+1}^{i=I_n} \lambda_i \right) + 2k(2I_n - k) \quad (15)$$

and the MDL criterion is given by

$$MDL(k) = -N \sum_{i=k+1}^{i=I_n} \log \lambda_i + N(I_n - k) \log \left(\frac{1}{I_n - k} \sum_{i=k+1}^{i=I_n} \lambda_i \right) + \frac{k}{2}(2I_n - k) \log N \quad (16)$$

where $(\lambda_i)_{1 \leq i \leq I_n}$ are I_n eigenvalues of the covariance matrix of the n -mode unfolding \mathcal{R} : $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{I_n}$, and N is the number of columns of the n -mode unfolding \mathcal{R} .

The n -mode rank K_n is the value of k ($k = 1, \dots, I_n - 1$) which minimizes AIC or MDL criterion.

Those criteria are integrated in the final algorithm summarized in the following section.

5 ALS Algorithm

An Alternative Least Square algorithm needs to be used to jointly find H_n n -mode Wiener filters that enables to reach the global minimum of mean square error $e(H_1, \dots, H_N)$ given by (3). One ALS algorithm can be summarized in the following steps:

1. initialization $k = 0$: $\mathcal{R}^0 = \mathcal{R} \Leftrightarrow H_n^0 = I_{I_n}$ for all $n = 1$ to N .
2. ALS loop: while $\|\mathcal{X} - \mathcal{R}^k\|^2 > \text{thr}$, with $\text{thr} > 0$ a priori fixed,
 - (a) for $n = 1$ to N :
 - i. $\mathcal{R}_n^k = \mathcal{R} \times_1 H_1^k \cdots \times_{n-1} H_{n-1}^k \times_{n+1} H_{n+1}^k \cdots \times_N H_N^k$,
 - ii. $H_n^{k+1} = \arg \min \| \mathcal{X} - \mathcal{R}_n^k \times_n Q_n \|^2$ subject to $Q_n \in \mathbb{R}^{I_n \times I_n}$.
 - (b) $\mathcal{R}^{k+1} = \mathcal{R} \times_1 H_1^{k+1} \cdots \times_N H_N^{k+1}$, $k \leftarrow k + 1$.
3. output: $\hat{\mathcal{X}} = \mathcal{R} \times_1 H_1^k \cdots \times_N H_N^k$.

Step (2)(a)(ii) of the ALS algorithm can be decomposed into the following sub-steps:

1. n -mode unfold \mathcal{R}_n^k into $R_n^k = R_n(H_1^k \otimes \cdots \otimes H_{n-1}^k \otimes H_{n+1}^k \cdots \otimes H_N^k)$, and \mathcal{R} into R_n ;
2. compute $\gamma_{RR}^n = E(R_n^k R_n^{kT})$, perform its EVD and place the eigenvalues in λ_k^γ , for $k = 1$ to I_n ;
3. estimate K_n using AIC (15) or MDL (16) criterion;
4. estimate $\sigma_\gamma^{(n)2}$ by computing $\frac{1}{I_n - K_n} \sum_{k=K_n+1}^{I_n} \lambda_k^\gamma$ and estimate β_i by computing $\lambda_i^\gamma - \sigma_\gamma^{(n)2}$, for $i = 1$ to K_n ;
5. compute $\Gamma_{RR}^{(n)} = E(R_n^k R_n^{kT})$, perform its EVD, keep in matrix V_s^n the K_n eigenvectors associated with the K_n largest eigenvalues of $\Gamma_{RR}^{(n)}$, and keep the R_n largest eigenvalues $\lambda_{I_k}^n$ for $k = 1$ to K_n ;
6. compute the $(k + 1)^{\text{th}}$ iteration of n -mode Wiener filter H_n^{k+1} using (13);

6 Performances and Hyperspectral Images

In this section, multiway Wiener filtering is applied and compared with classical signal subspace based methods to improve the SNR of hyperspectral images. The first classical bidimensional processing methods basically consist in a consecutive

Wiener filtering of each two-dimensional spectral channel. The second method consists to a preprocessing by projection on the spectral mode to decorrelate the different channels each other, then Wiener filtering is applied on each two-dimensional spectral channel. In both applications, the efficiency of denoising is tested in presence of additive white Gaussian noise.

This noise, \mathcal{N} , can be modeled by

$$\mathcal{N} = \alpha \cdot \mathcal{G} \quad (17)$$

in which every element of $\mathcal{G} \in \mathbb{R}^{I_1 \times I_2 \times I_3}$ is an independent realization of a normalized centered Gaussian law, and where α is a coefficient that enables to set the SNR in noisy data tensor \mathcal{G} .

Let's define the Signal to Noise Ratio (SNR in dB) in the noisy data tensor.

$$SNR = 10 \times \log\left(\frac{\|\mathcal{X}\|^2}{\|\mathcal{B}\|^2}\right) \quad (18)$$

In order to a posteriori verify the quality of estimation of the different tensor filtering, it is possible to use the Relative Reconstruction Error (RRE) defined as follows:

$$RRE = \frac{\|\hat{\mathcal{X}} - \mathcal{X}\|^2}{\|\mathcal{X}\|^2}. \quad (19)$$

The RRE criterion enables a qualitative comparison between multiway Wiener and classical filtering.

The hyperspectral images with 300 rows, 300 columns and 146 spectral channels (from 435 nm to 2326 nm) are considered initially. It can be modeled by tensor $\mathcal{X} \in \mathbb{R}^{300 \times 300 \times 146}$. For example Figure 4 shows image of channel 20.

Experiment 1: Detection criteria:

We have applied the proposed criteria on several noisy tensors $\mathcal{X} \in \mathbb{R}^{300 \times 300 \times 20}$ constructed from the twenty first channels and white Gaussian tensor noise. We have compared the optimal n -mode tensor rank estimated thanks to the lower rank- (K_1, K_2, K_3) tensor approximation, LRTA (K_1, K_2, K_3) [9] to the estimated values by the proposed criteria. In this study, the optimal estimation is in the sense of minimization of the RRE criterion defined above. Figure 3(a) shows the results obtained for 1-mode rank estimation. One can see that AIC criterion and LRTA give the same results for all SNR values. Indeed, the optimal 1-mode rank increases when the SNR value increases. Figure 3(b) shows the two criteria efficiency when the LRTA is used with the optimal n -mode ranks given by AIC or MDL criterion.

Experiment 2: Multiway Wiener filtering:

In order to evaluate the performances of the proposed multiway filtering method, some signal-independent white Gaussian noise \mathcal{N} , is added to \mathcal{X} and results in noisy image $\mathcal{R} = \mathcal{X} + \mathcal{N}$. Channel 20 of noisy hyperspectral image \mathcal{R} is represented, for example, on Figure 4, and corresponds to a global computed on tensor \mathcal{R} of 17.3 dB.

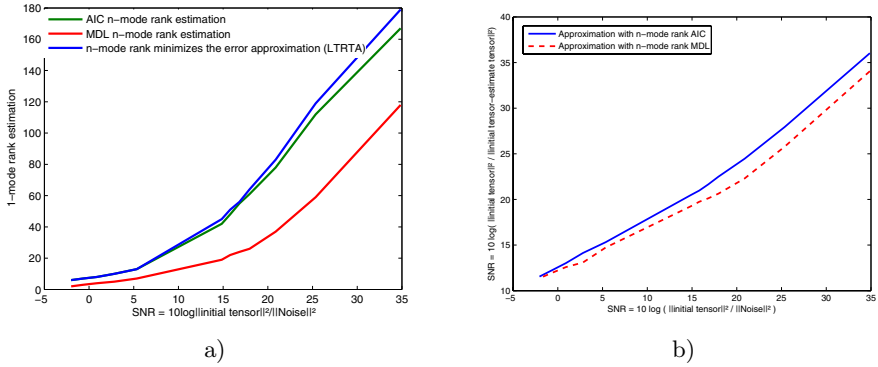


Fig. 3. n -mode rank estimation of a noisy tensor \mathcal{X} . a) AIC, MDL and LRTA(k_1, k_2, k_3) n -mode rank estimation of a noisy tensor. b) Approximation results with the two-criteria AIC and MDL function of SNR.

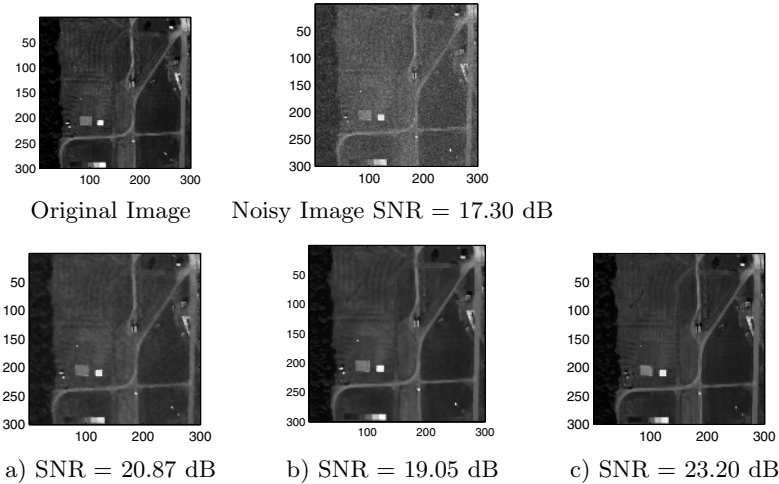


Fig. 4. Different methods of denoising: a) Image denoised with 2D Wiener channel by channel SNR = 20.87 dB. b) Image denoised first whitening on the spectral mode then 2D Wiener SNR = 19.05 dB. c) Image denoised with Wiener multiway SNR = 23.20 dB.

Figures 4 (a), 5 (a) and 6(a) represent channel 20 of the hyperspectral image obtained by applying channel-by-channel-based wiener-filtering on noisy image \mathcal{R} . Figures 4(b), 5(b) and 6(b) represent channel 20 of the hyperspectral image obtained by applying channel-by-channel-based wiener-filtering on noisy image \mathcal{R} after whitening the data in spectral mode.

Finally, Figures 4(c), 5(c) and 6(c) represent channel 20 of the hyperspectral image obtained by applying the proposed multiway wiener filtering on noisy image \mathcal{R} . According to the last simulation, the proposed method gives better

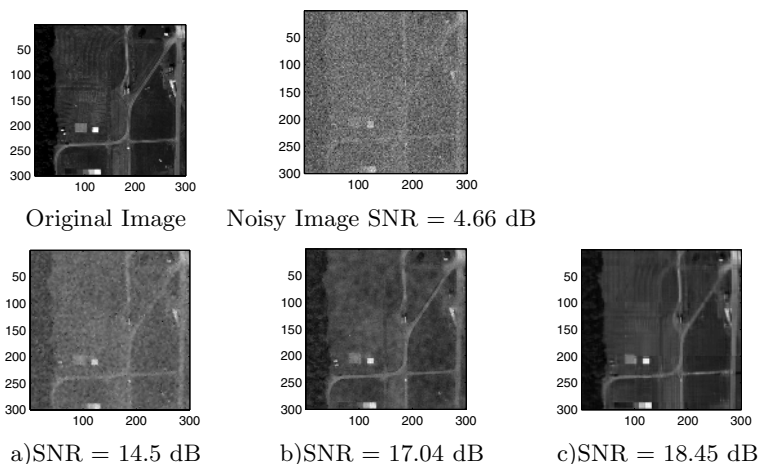


Fig. 5. Different method of denoising: a) Image denoised with 2D Wiener channel per channel SNR = 14.5 dB. b) Image denoised first whitening on the spectral mode then 2D Wiener SNR = 17.04 dB. c) Image denoised with Wiener multiway SNR = 18.45 dB.

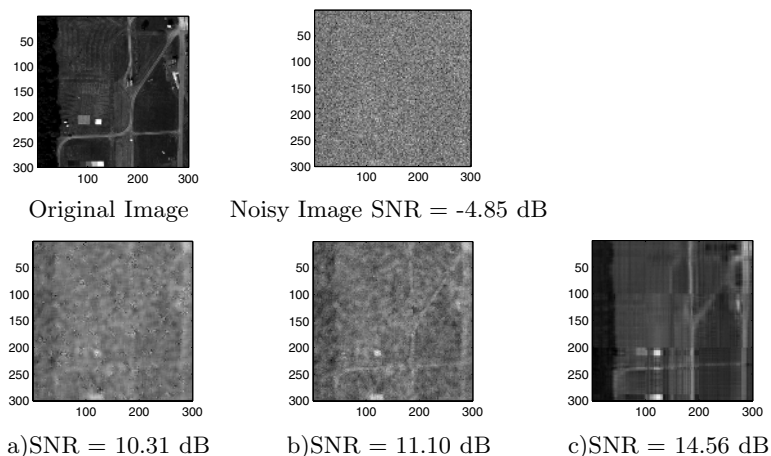


Fig. 6. Different methods of denoising: a) Image denoised with 2D Wiener channel per channel SNR = 10.31 dB. b) Image denoised first whitening on the spectral mode then 2D Wiener SNR = 11.10 dB. c) Image denoised with Wiener multiway SNR = 14.56 dB.

results than channel-by-channel wiener filtering in regard to the improvement of SNR.

Moreover, the evolution of the RRE with respect to the SNR varying from -5 dB to 30 dB, represented on Figure 7 shows that the RRE obtained with the proposed method is lower than the one obtained with previously existing methods.

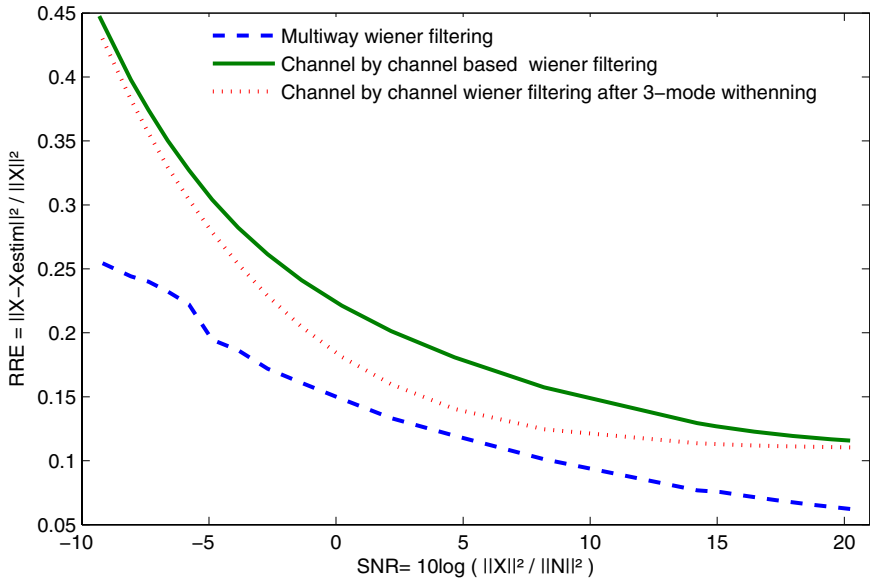


Fig. 7. Performance of multiway wiener filtering in the case of noise reduction in hyperspectral images \mathcal{X}

In each of this situation, the multiway Wiener filtering improves considerably the SNR.

7 Conclusion

In this paper we have described a new algorithm for denoising a tensor of arbitrary order. For hyperspectral images, we have proposed a tensor model to consider all data as whole tensor. The proposed multiway filtering is an extension of bidimensional wiener filtering to tensor signal. In order to estimate the signal subspace for each mode we have extended the well-known criteria AIC and MDL to tensor signal. Since filters that minimize the mean squared error need to be determined simultaneously, an ALS algorithm has been developed : both spatial and spectral informations are taken into account conjointly . A simulation involving several tensors with known rank for each mode shows the efficiency of the proposed criteria. It was also confirmed that the multiway filtering improves significantly more efficiency the SNR than the classical methods in several experiments with hyperspectral images.

References

1. Comon, P.: Tensor decompositions, state of the art and applications. In: IMA Conf. mathematics in Signal Processing, Warwick, UK (2000)
2. Muti, D., Bourennane, S.: Multidimensional filtering based on a tensor approach. Signal Processing Journal, Elsevier **85** (2005) 2338–2353

3. Muti, D., Bourennane, S.: Multiway filtering based on fourth order cumulants. *Applied Signal Processing, EURASIP* **7** (2005) 1147–1159
4. Tucker, L.: Some mathematical notes on three-mode factor analysis. *Psychometrika* **31** (1966) 279–311
5. Muti, D., Bourennane, S.: Survey on tensor signal algebraic filtering. *Signal Processing Journal, Elsevier*. To be published (2006)
6. Le Bihan, N.: Traitement algébrique des signaux vectoriels: Application à la séparation d'ondes sismiques. Phd thesis, INPG, Grenoble, France (2001)
7. Muti, D., Bourennane, S.: Multidimensional signal processing using lower rank tensor approximation. In: *IEEE Int. Conf. on Acoustics, Systems and Signal Processing*, Hong Kong, China (2003)
8. Wax, M., Kailath, T.: Detection of signals information theoretic criteria. *IEEE International Conference on Acoustics Speech and Signal Processing* **33** (1985) 387–392
9. De Lathauwer, L., De Moor, B., Vandewalle, J.: A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications* **21** (2000) 1253–1278

A Linear-Time Approach for Image Segmentation Using Graph-Cut Measures^{*}

Alexandre X. Falcão, Paulo A.V. Miranda, and Anderson Rocha

Institute of Computing – State University of Campinas (UNICAMP)
CEP 13084-851, Campinas, SP – Brazil
{afalcao, paulo.miranda, anderson.rocha}@ic.unicamp.br

Abstract. Image segmentation using graph cuts have become very popular in the last years. These methods are computationally expensive, even with hard constraints (seed pixels). We present a solution that runs in time proportional to the number of pixels. Our method computes an ordered region growing from a set of seeds inside the object, where the *propagation order* of each pixel is proportional to the cost of an *optimum path* in the image graph from the seed set to that pixel. Each pixel defines a region which includes it and all pixels with lower propagation order. The boundary of each region is a possible cut boundary, whose cut measure is also computed and assigned to the corresponding pixel on-the-fly. The object is obtained by selecting the pixel with minimum-cut measure and all pixels within its respective cut boundary. Approaches for graph-cut segmentation usually assume that the desired cut is a global minimum. We show that this can be only verified within a reduced search space under certain hard constraints. We present and evaluate our method with three cut measures: normalized cut, mean cut and an energy function.

1 Introduction

We consider the problem of segmenting an image in object and background by graph-cut measures. The image is interpreted as an undirected graph, whose nodes are the image pixels and whose arcs are weighted and defined by an adjacency relation between pixels. We wish to assign weights to the arcs and define an objective function (a graph-cut measure), such that its minimum corresponds to the desired segmentation (i.e., a *cut boundary* whose arcs connect the nodes between object and background).

Approaches for graph-cut segmentation usually aim at assigning higher weights to arcs inside object and background, and lower weights otherwise. Their objective functions measure some global property of the object's boundary from this weight assignment. Wu and Leahy [1] were the first to introduce a solution for graph cut using as measure the sum of the arc weights in the cut boundary.

^{*} The authors thank the financial support of FAPESP (Procs. 03/13424-1, 05/58103-3, and 05/59808-0) and CNPq (Proc. 302427/04-0).

Their cut measure has the bias toward small boundaries and other objective functions, such as *average cut* [2], *mean cut* [3], *average association* [4], *normalized cut* [5], *ratio cut* [6], and *energy functions* [7, 8, 9] have been proposed to circumvent this problem.

The problem of finding a minimum of an objective function through graph cut is NP-hard for a generic graph and very often solutions require hard constraints. Heuristic solutions have been proposed in polynomial time [10], but with poor computational performance, and the results are sometimes far from the desired segmentation [11]. Indeed we have verified that even in a reduced search space that includes the desired cut, it does not always correspond to the minimum cut. This suggests that hard constraints are really needed in practice. For example, two terminal nodes (*source* and *sink*) can be added to the image graph, representing object and background respectively [7, 8]. Additionally to the weight assignment between pixels, this approach aims at assigning lower arc-weights between source and object pixels, higher arc-weights between sink and object pixels, lower arc-weights between sink and background pixels, and higher arc-weights between source and background pixels. A min-cut/max-flow algorithm from source to sink [12, 13] is used to compute the minimum-cut boundary. If the method fails the detection of the desired boundary, the user can impose the arc weights with source and sink by selecting seed pixels inside and outside the object [7]. The running time of these algorithms is still polynomial [8] (i.e., typically $O(mn^2)$ where m is the number of arcs and n is the number of *nodes*).

We present a solution that runs in linear time (i.e., in $O(n)$). Our method computes an ordered region growing from a set of seeds inside the object, where the *propagation order* of each pixel is proportional to the cost of an *optimum path* in the image graph from the seed set to that pixel. Each pixel defines a region which includes it and all pixels with lower propagation order. The boundary of each region is a possible cut boundary, whose cut measure is also computed and assigned to the corresponding pixel on-the-fly. The object is obtained by selecting the pixel with minimum-cut measure and all pixels within its respective cut boundary.

Our method essentially reduces the search space by ordering possible cuts from inside to outside the object. It requires lower arc weights across the object's boundary than inside it in order to include the desired cut in the reduced space. When this weight assignment is not achieved, the method can still work by adding more seeds. A problem, however, has been the sensitivity of some cut measures with respect to the heterogeneity (arc weights) outside the object. We evaluate this aspect with normalized cut [5], mean cut [3], and an energy function [7, 9].

We could use the same adjacency relation, weight assignment between pixels, and energy function to compare our method with the one by Boykov and Jolly [7] in the context of interactive segmentation. However, under the same conditions, both methods are likely to produce similar results except to the fact that our algorithm is more efficient. Instead of that, we prefer to verify the accuracy of our approach in a real application that represents the worst case for the aforementioned cut measures.

Section 2 presents the image graph, weight functions, and cut measures used in this paper. We present our method for the 2D case, but its extension to 3D is straightforward. The method and its algorithm are presented in Section 3. Section 4 evaluates it using three cut measures and our conclusions are stated in Section 5.

2 Image Graphs and Cut Measures

Consider an undirected graph where the pixels are the nodes and the arcs are defined by an irreflexive 4-adjacency relation between pixels. There are many ways of exploiting image features to compute arc weights [5, 7, 14]. We suggest to assign a membership value for each pixel with respect to the object based on image features (texture, color, gradients), which may be different depending on the application. The idea is to improve the weight assignment by reducing inhomogeneities inside the object.

Let \mathbf{x}_p be a feature vector computed at a given pixel p ; μ_p and Σ_p be mean and covariance matrices of the feature vectors \mathbf{x}_q computed at all pixels q within an adjacency radius around p ; and T be a set of training pixels, selected in object regions that have different image features. For a given pixel $s \in T$, we compute a membership value $R_s(p)$ for every image pixel p .

$$R_s(p) = \exp\left(-\frac{1}{2d}(\mathbf{x}_p - \mu_s)^t \Sigma_s^{-1}(\mathbf{x}_p - \mu_s)\right) \quad (1)$$

where $d > 1$ takes into account the absence of statistical information (e.g., we use $d = 10$). We also set a distinct adjacency radius for each pixel $s \in T$, making it as largest as possible, in order to compute the best estimation for μ_s and Σ_s inside the object region that includes s . A region map R is obtained as

$$R(p) = \max_{\forall s \in T} \{R_s(p)\}. \quad (2)$$

We also apply a median filter on R to make it more homogeneous. The weight $w(p, q)$ for any arc (p, q) is given by

$$w(p, q) = \exp\left(-\frac{(R(p) - R(q))^2}{2d}\right). \quad (3)$$

Figures 1a–c show three original images, where the training pixels and their adjacency radii are indicated by circles. The respective region maps are shown in Figures 1d–f. We used two normalized attributes within $[0, 1]$ for the feature vectors of Equation 1 in each case: brightness and gradient magnitude (Figure 1a); and red and green values (Figures 1b and 1c). Note that the choice of these attributes is a separate problem, and the segmentation can not be generally solved by thresholding the region map and extracting the binary components, which are hard-connected to internal seeds (e.g., Figure 1d).

Due to the heterogeneity of the background, it is very difficult to obtain higher arc weights outside the object. This affects some graph-cut measures more than

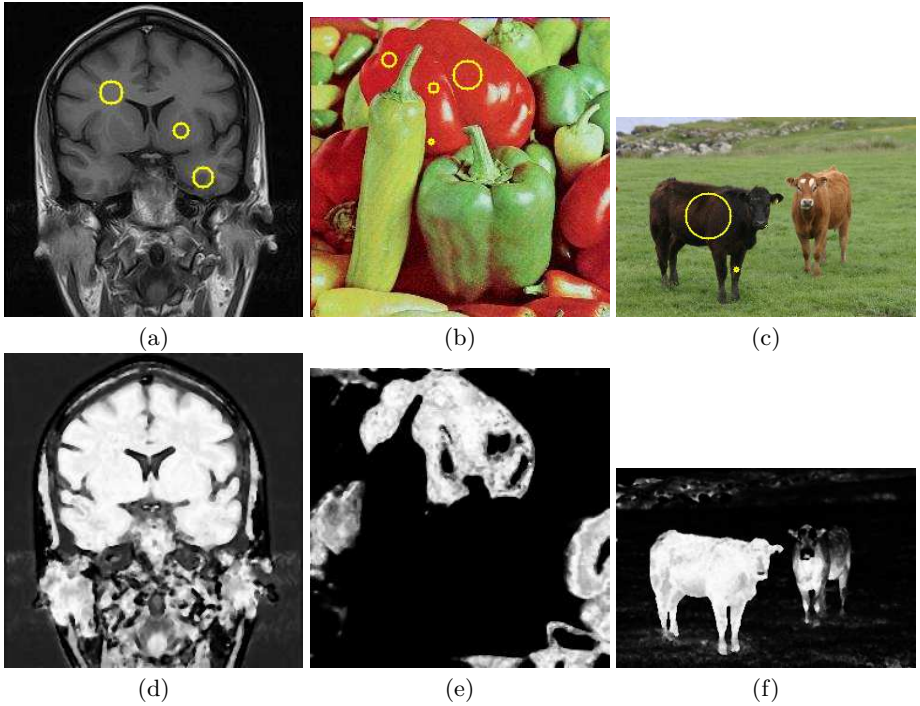


Fig. 1. (a) A Magnetic Resonance (MR) image of a brain with three training pixels (the circles indicate their adjacency radii). (b) A colored image of peppers with four training pixels. (c) A colored image of two cows with two training pixels. (d–f) The respective region maps of (a),(b), and (c).

others. Therefore, we will consider the normalized cut [5], mean cut [3], and an energy function [7, 9] to evaluate this aspect in Section 4.

Let I and E be the interior and exterior of a cut boundary IE , which consists of a set of arcs (p, q) where $p \in I$ and $q \in E$. The normalized cut is defined as

$$\frac{cut(I, E)}{asso(I) + cut(I, E)} + \frac{cut(I, E)}{asso(E) + cut(I, E)} \quad (4)$$

where

$$cut(I, E) = \sum_{\forall (p, q) | p \in I, q \in E} w(p, q) \quad (5)$$

$$asso(I) = \sum_{\forall (p, q) | p \in I, q \in I} w(p, q) \quad (6)$$

$$asso(E) = \sum_{\forall (p, q) | p \in E, q \in E} w(p, q). \quad (7)$$

The mean cut is defined as

$$\frac{cut(I, E)}{|IE|} \quad (8)$$

where $|IE|$ is the number of arcs in IE .

We have chosen an energy function similar to that proposed in [7] and consistent with the general formulation described in [9].

$$\lambda \left(\sum_{\forall p \in I} (1 - Ro(p)) + \sum_{\forall q \in E} (1 - Rb(q)) \right) + cut(I, E) \quad (9)$$

where Ro and Rb are region maps computed by Equation 2 using training pixels inside object and background, respectively; and $\lambda > 0$ represents the importance of the first term (a normalization factor) with respect to the second one.

3 Region Growing by Ordered Propagation with Graph Cut

Let $A_4(p)$ be the set of the 4-adjacent pixels of p , excluding it. A *path* π in the image graph (Section 2) is a sequence $\langle p_1, p_2, \dots, p_n \rangle$, such that $p_{i+1} \in A_4(p_i)$, for $i = 1, 2, \dots, n - 1$.

First, assume that Equation 3 assigns lower arc weights across the object's boundary than inside it. These arc weights are inversely proportional to the dissimilarities $\delta(p, q)$ between 4-adjacent pixels of the region map. For a given set S of internal seeds, we define the *cost* c of a path π as:

$$c(\pi) = \begin{cases} \max_{i=1,2,\dots,n-1} \{\delta(p_i, p_{i+1})\} & \text{if } p_1 \in S \\ +\infty & \text{otherwise} \end{cases} \quad (10)$$

where $\delta(p, q) = K(1 - w(p, q))$ for an integer K that represents the maximum dissimilarity between pixels (e.g., $K = 1023$). The reason for using an integer K will be explained later.

A path from a seed set S to a pixel p is *optimum* when its cost is minimum as compared to the cost of any other path from S to p . Under the above conditions, it is enough to have a single seed in S (we will discuss later the case of multiple seeds), and the optimum paths from S to object pixels will have costs strictly less than the costs of optimum paths with terminus at background pixels. The object could be detected by thresholding the costs of the optimum paths from S , but this threshold is unknown. Thus, we grow a region from S by aggregating one adjacent pixel at a time in order proportional to the cost of an optimum path from S to that pixel; such that the object pixels will be aggregated before the background pixels.

Each pixel defines a region which includes it and all pixels with lower propagation order. The boundary of each region is a possible cut boundary, whose cut measure is also computed and assigned to the corresponding pixel on-the-fly. The desired cut boundary consists of arcs between object and background pixels,

and the object is defined by the pixel with minimum-cut measure and all pixels within its respective cut boundary.

This region growing process creates a reduced search space that includes the desired cut boundary. Now it is expected that the objective function be able to detect it as the one with minimum cut. This is certainly not a problem when Equation 3 assigns lower arc weights across the object’s boundary than inside **and outside** it.

If Equation 3 assigns low arc weights inside the object, the method may require one seed for each part of the object that satisfies the above conditions. The cut boundaries from each seed will merge into the desired cut boundary before the optimum paths reach the background pixels.

3.1 Algorithm

Our method uses the *Image Foresting Transform* (IFT)— a tool for the design of image processing operators based on connectivity [15]. The IFT algorithm essentially reproduces the aforementioned process by assigning an optimum path from S to every pixel in a non-decreasing order of cost. Its bottleneck is a priority queue Q , which selects a path of minimum cost $C(p)$ at each iteration by removing its last pixel p from Q . Ties are broken in Q using first-in-first-out policy. The algorithm runs in linear time if $\delta(p, q)$ is an integer in $[0, K]$ and Q is implemented as described in [16].

We need to modify the IFT algorithm as follows. When a pixel p is removed from Q , p receives a propagation order $O_d(p) \in [1, n]$, for an image with n pixels. At this moment, p and all pixels with lower propagation order define a region I and the algorithm has found the optimum paths from S to every pixel in I [15]. The remaining pixels define a region E ; the cut IE is defined by arcs between pixels of I and its 4-adjacent pixels in Q ; and the cut measure $M(p)$ for IE is computed on-the-fly. We first illustrate these modifications for normalized cut.

Algorithm 1 COMPUTATION OF THE PROPAGATION ORDER MAP O_d AND NORMALIZED CUT MAP M

INPUT: An image and adjacency A_4 .

OUTPUT: Maps O_d and M .

AUXILIARY: A priority queue Q and variables o , ai , ie , and ae that store the order and values of the Equations 5- 7 for the cut IE .

1. Set $o \leftarrow 1$, $ai \leftarrow 0$, $ie \leftarrow 0$, and $ae \leftarrow 0$.
2. **For** every image pixel p , **do**
3. Set $C(p) \leftarrow +\infty$ and $O_d(p) \leftarrow +\infty$.
4. **For** every pixel $q \in A_4(p)$ **do**
5. L Set $ae \leftarrow ae + w(p, q)/2$.
6. **For** every pixel $p \in S$ **do**
7. L Set $C(p) \leftarrow 0$ and insert p in Q .
8. **While** Q is not empty **do**
9. | Remove p from Q such that $C(p)$ is minimum.
10. | **For** every pixel $q \in A_4(p)$ **do**

```

11.   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
12.   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
13.   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
14.   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
15.   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
16.   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
17.   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
18.   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
19.   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
20.   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |
21.   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |   |

```

If $O_d(q) < O_d(p)$, **then**
 \perp *Set* $ie \leftarrow ie - w(p, q)$ and $ai \leftarrow ai + w(p, q)$.
Else
 Set $ie \leftarrow ie + w(p, q)$ and $ae \leftarrow ae - w(p, q)$.
 Set $cst \leftarrow \max\{C(p), \delta(p, q)\}$.
 If $cst < C(q)$ **then**
 If $C(q) \neq +\infty$ **then**
 \perp *Remove* q from Q .
 \perp *Set* $C(q) \leftarrow cst$ and *insert* q in Q .
 Set $O_d(p) \leftarrow o$ and $o \leftarrow o + 1$.
 \perp *Set* $M(p) \leftarrow \frac{ie}{ie + ai} + \frac{ie}{ie + ae}$.

Lines 1–7 initialize maps, variables and insert seed pixels in Q . The division by 2 in Line 5 takes into account that the graph is undirected (i.e., $w(p, q) = w(q, p)$ should be considered only once). Thus, variable ae is initialized with the sum of all arc weights in the graph. Lines 8–21 compute the maps M and O_d during the IFT. When p is removed from Q (line 9), it leaves E and goes to I . At this moment, all arcs that contain p need to be evaluated. The condition stated in Line 11 indicates that $q \in I$, then arc (p, q) is being removed from IE and its weight must be considered to update ie and ai . Otherwise $q \in E$, then arc (p, q) is being inserted in IE and its weight must be used to update ie and ae . Lines 15–19 evaluate if the path that reaches q through p is better than the current path with terminus q and update Q and $C(q)$ accordingly. Finally, lines 20–21 compute the propagation order of p and the measure of its corresponding cut IE . After Algorithm 1, the object is obtained by selecting a pixel m with minimum-cut measure and thresholding O_d at values less than or equal to $O_d(m)$.

The above algorithm can be easily modified for mean cut if we set a variable n_{ie} to 0 in line 1 (where n_{ie} stores the size of IE); compute ie as above; insert $n_{ie} \leftarrow n_{ie} - 1$ in line 12 and $n_{ie} \leftarrow n_{ie} + 1$ in line 14; and set $M(p)$ to ie/n_{ie} in line 21. In the case of the energy function, we substitute lines 4 and 5 by $ae \leftarrow ae + (1 - R_b(p))$; compute ie as above; remove the computation of ai and ae from lines 12 and 14; and insert $ai \leftarrow ai + (1 - R_o(p))$ and $ae \leftarrow ae - (1 - R_b(p))$ between lines 20 and 21. In line 21, we set $M(p)$ to $\lambda(ai + ae) + ie$. Note that, we can do the same for many other graph-cut measures (e.g., [2, 4, 9]).

4 Results and Evaluation

Figures 1d and 1e show that the cut boundary may contain multiple contours due to “holes” (dark regions) inside the region map. The holes may be part of the object (Figure 1e) or not (Figure 1d). This problem may occur in any graph-cut segmentation approach. In our method, we close the holes in the resulting binary image and consider only the external contour as object boundary. Some results using the region maps of Figure 1 are presented in Figure 2 for normalized cut, mean cut, and energy function. In the latter, we also used training pixels outside the object to compute the background region map R_b of Equation 9.

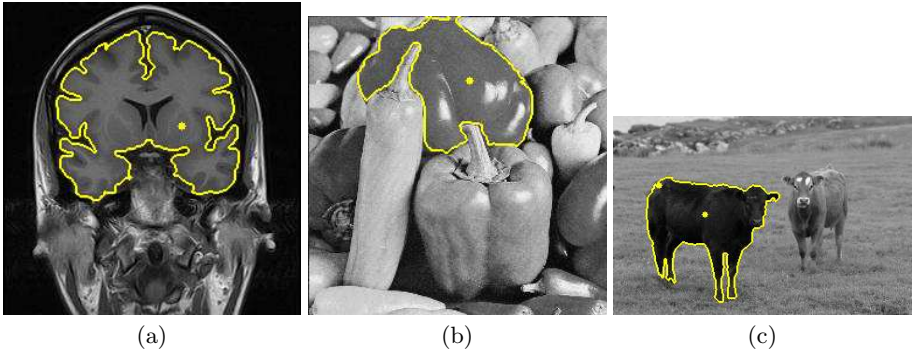


Fig. 2. Segmentation results where the seeds are indicated by dots, using (a) normalized cut, (b) mean cut, and (c) energy function

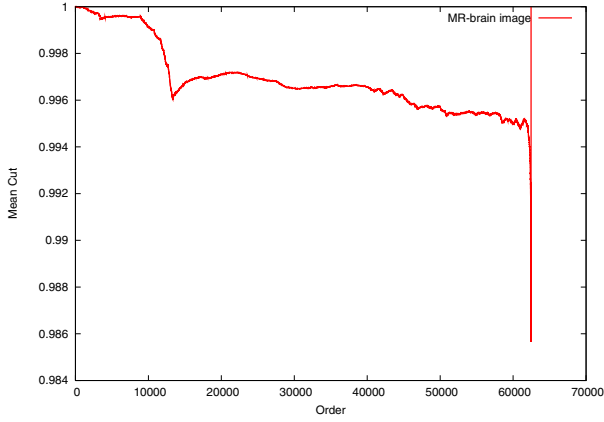
Figure 3 shows the cut measure versus the pixel propagation order for mean cut, normalized cut, and energy function using the region map of Figure 1d. In the case of the energy function, we also created a background map and set λ to 80 in Equation 9. In all cases the IFT parameters are the same and the desired cut occurs at order 13,340 of the reduced search space. However, it corresponds to the minimum cut only for the energy function (Figure 3c). Mean cut and normalized cut fail because of the weight assignment outside the object (Figures 3a–b). On the other hand, both cut measures can work if we add more hard constraints, such as limiting the search up to some propagation order o , for $o < n$ and greater than the object’s size (e.g., $o = 0.7n$ in this case).

This shows that any approach to separate object and background using graph cut is likely to require some hard constraints, because the problem can not be simply reduced to finding a minimum of an objective function in the entire search space. Since false-cut boundaries due to similarities between object and background are very common in practice, we have chosen a real application representing the worst case in this respect to evaluate our method.

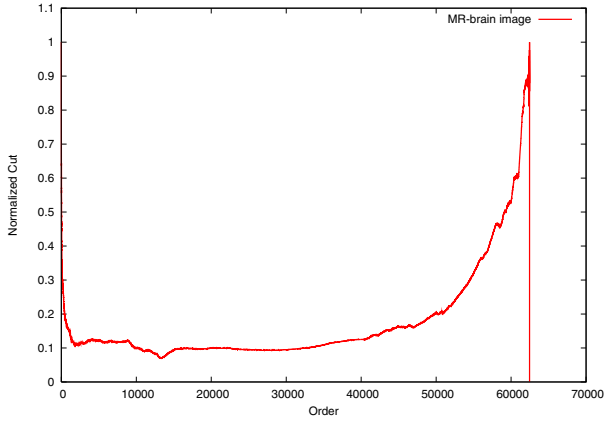
4.1 Experiments for Evaluation

We have selected 6 images of archaeological fragments, similar to the one shown in Figure 4a. In this application, the boundary of each fragment has to be perfectly detected to reassemble the original object [17]. Thus, any failure in the detected boundary is considered a segmentation error. The similarities between object and background and touching fragments fail segmentation by thresholding.

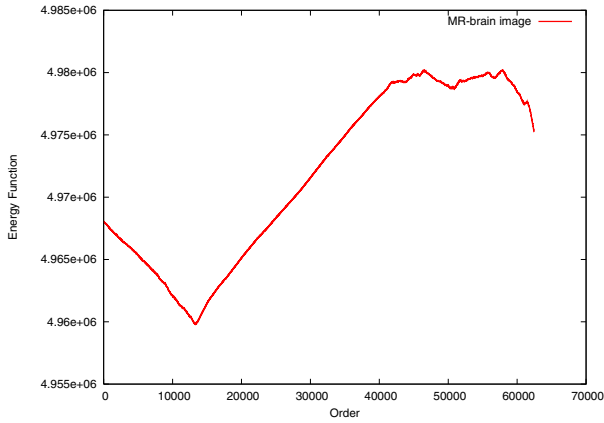
The images have 512×384 pixels ($n = 196,608$) and a total of 211 fragments. We applied morphological operations to reduce internal noise, eliminate the grid pattern in the background, and estimate one seed pixel inside each fragment. This approach was able to find seeds inside 201 out of the 211 fragments automatically. Therefore, our experiments consist of using the method to detect the boundary of 201 seeded fragments in the filtered images.



(a)



(b)



(c)

Fig. 3. The cut measure versus the pixel propagation order for (a) mean cut, (b) normalized cut, and (c) energy function using the MR-brain image

A suitable region map for each fragment would require seed selection on the shadow region that appears on most fragments. Since this is impractical in an automatic fashion, we decided to use dissimilarity and weight functions based on differences of brightness, as usually done in graph-cut segmentation [5, 3, 7].

$$\delta(p, q) = |f(p) - f(q)| \quad (11)$$

$$w(p, q) = 1.0 - \frac{\delta(p, q)}{K} \quad (12)$$

where $f(p)$ is the brightness of pixel p and K is the maximum brightness value in the filtered image. However, the region maps Ro and Rb were computed for the entire image (taking into account that fragments and non-fragments have dissimilar features) and used in the following energy function.

$$\lambda \left(\sum_{\forall p \in I} (1 - Ro(p)) + \sum_{\forall q \in E} (1 - Rb(q)) \right) + \sum_{\forall (p, q) | p \in I, q \in E} \alpha(p, q) w(p, q) \quad (13)$$

where

$$\alpha(p, q) = \begin{cases} 0 & \text{if } Ro(p) > Rb(p) \text{ and} \\ & Ro(q) < Rb(q) \\ 1 & \text{otherwise.} \end{cases} \quad (14)$$

and $\lambda = 40$. Note that Equation 14 uses Ro and Rb to restrict the computation of $w(p, q)$ inside uncertainty regions, as suggested in [7].

Our strategy is to assign a distinct number for each seed, detect each fragment separately, and label it with its corresponding number (see examples in Figures 4b–c). Some fragments touch each other, but the algorithm can separate them. When the algorithm fails, it usually outputs the union of two touching fragments twice, one for each seed. This situation is automatically detected and the fragments are separated by watershed transform restricted to their union [15].

The method with normalized cut correctly detected only 52 fragments (25.87%). In order to confirm that this bad result was not due to the IFT, we repeated the experiment with normalized cut and mean cut, but we limited the search for the minimum-cut value up to order $o = 0.05n$. The method with normalized cut correctly detected 104 (51.74%) fragments, while the method with mean cut detected 190 (94.53%) fragments correctly.

We also performed the experiments with the energy function. In this case, the method correctly detected 182 (90.50%) fragments. Although the number of correct detections was lower than using mean cut with $o = 0.05n$, we have observed that energy functions are usually more robust than the other two cut measures, when it is possible to devise a suitable normalization factor in Equation 9.

Finally, the mean running time to execute the method over images with 512×384 pixels was 161 milliseconds, using a 2.8GHz Pentium IV PC.

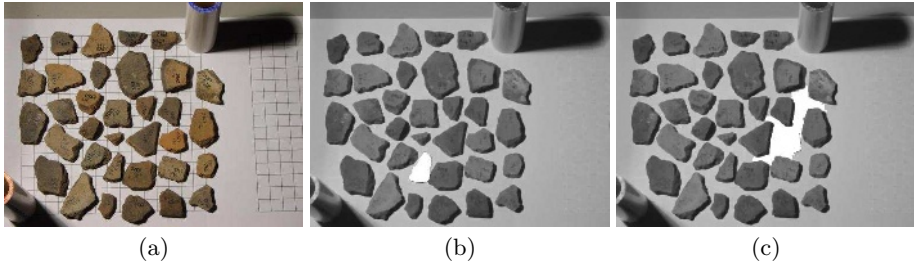


Fig. 4. Detection of archaeological fragments. (a) the original image. (b-c) Examples of correct and incorrect detections.

5 Conclusion

We showed that the segmentation by graph cut usually requires hard constraints to find the desired cut as minimum cut. We proposed a linear-time solution where the desired cut is included in a reduced search space under certain hard constraints applied to arc weight assignment and seed selection. We presented and evaluated our method for three cut measures: normalized cut, mean cut and an energy function.

The method requires proper weight assignment and/or more seeds inside the object, such that the IFT can reproduce its boundary during the region growing process. Under this condition, the problem is reduced to the sensitivity of the cut measures with respect to the weight assignment outside the object. The experiments evaluated this aspect in the worst case (i.e., when object and background parts have similar image properties). Even so, the results show accuracy greater than 90% for some cut measures. Therefore, we may conclude that our approach is a significant contribution in graph-cut segmentation.

In interactive segmentation, the IFT allows competition among internal and external seeds [18]. The combination of external seeds (to reduce the search space) and cut measures (to reduce user intervention) may provide more effective solutions than using [18, 7]. We are currently investigating this variant.

References

1. Wu, Z., Leahy, R.: An optimal graph theoretic approach to data clustering: theory and its applications to image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **15** (1993) 1101–1113
2. Cox, I.J., Rao, S.B., Zhong, Y.: Ratio regions: a technique for image segmentation. In: *Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*. (1996) 557–564
3. Wang, S., Siskind, J.M.: Image segmentation with minimum mean cut. In: *Intl. Conf. on Computer Vision (ICCV)*. Volume 1. (2001) 517–525
4. Sarkar, S., Boyer, K.L.: Quantitative measures of change based on feature organization: eigenvalues and eigenvectors. In: *Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*. (1996) 478–483

5. Shi, J., Malik, J.: Normalized cuts and image segmentation. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **22** (2000) 888–905
6. Wang, S., Sinkind, J.M.: Image segmentation with ratio cut. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **25** (2003) 675–690
7. Yuri Y. Boykov, M.P.J.: Interactive graph cuts for optimal boundary & region segmentation of objects in N-D images. In: *Intl. Conf. on Computer Vision (ICCV)*. Volume 1. (2001) 105–112
8. Boykov, Y., Kolmogorov, V.: An experimental comparison of min-cut/max-flow algorithms for energy minimization in vision. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **26** (2004) 1124–1137
9. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **26** (2004) 147–159
10. Fowlkes, C., Belongie, S., Malik, J.: Efficient spatiotemporal grouping using the nystrom method. In: *Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*. (2001) 231–238
11. Carballido-Gamio, J., Belongie, S.J., Majumdar, S.: Normalized cuts in 3D for spinal MRI segmentation. *IEEE Trans. on Medical Imaging* **23** (2004) 36–44
12. Ford, L., Fulkerson, D.: *Flows in networks*. Princeton University Press (1962)
13. Greig, D., Porteous, B., Seheult, A.: Exact maximum a posteriori estimation for binary images. *J. Royal Statistical Society, series B* **51** (1989) 271–279
14. Kohli, P., Torr, P.H.: Efficiently solving dynamic markov random fields using graph cuts. In: *Intl. Conf. on Computer Vision (ICCV)*. Volume II. (2005) 922–929
15. Falcão, A., Stolfi, J., Lotufo, R.: The image foresting transform: Theory, algorithms, and applications. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **26** (2004) 19–29
16. Falcão, A., Udupa, J., Miyazawa, F.: An ultra-fast user-steered image segmentation paradigm: Live-wire-on-the-fly. *IEEE Trans. on Medical Imaging* **19** (2000) 55–62
17. Leitao, H., Stolfi, J.: A multiscale method for the reassembly of two-dimensional fragmented objects. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **24** (2002) 1239–1251
18. Falcão, A., Bergo, F.: Interactive volume segmentation with differential image foresting transforms. *IEEE Trans. on Medical Imaging* **23** (2004) 1100–1108

The RIM Framework for Image Processing

Øyvind Ryan*

Department of Informatics, Group for Digital Signal Processing and Image Analysis,
University of Oslo, P.O. Box 1080 Blindern, NO-0316 Oslo, Norway
oyvindry@ifi.uio.no

Abstract. A new design for image processing frameworks is proposed. The new design addresses high-level abstractions suited for component-based image processing applications, in particular real-time image processing with high performance demands. The RIM framework, an implementation of this design, is gone through. It is explained how RIM can be adapted in applications, and integrated with other image libraries. It is also shown how it can be used to confirm some properties of widely used image formats.

1 Introduction

This paper studies a recently developed image processing framework. The framework is called the Raster Imaging Framework, or RIM. Focus will be on what may be called *dynamic* image processing frameworks, since RIM can be placed in this category. A dynamic image processing framework doesn't concern itself with storing results persistently. Rather it is concerned with delivering *ephemeral images*, which may be based on an image composition description. The result is kept in memory until some other software has made use of it. Other software may be a web server transmitting the result to a client, a program storing the result to file or a graphical user interface displaying the result on screen. Frequent requests are typical, so memory usage and performance are important factors.

The paper [1] discusses in detail performance of some parts of RIM. To study RIM's dynamic image processing capabilities closer, the concept of *lazy evaluation* was introduced. Lazy evaluation means to process ephemeral images piece by piece, such as scanline by scanline, keeping only small parts of the image in memory at a time. This reduces the working set [2], [3] of the image processing.

There is an application-driven need for dynamic image processing libraries. A typical application is to extract a small section of a large image, and convert it to another image format. Such applications often come in the form of requests to a server, in particular a *map server*. The OpenGIS consortium has established a standard for map servers, called *WMS*, or *Web Map Server* [4]. WMS specifies the behaviour of a service which produces georeferenced maps.

* This project has been sponsored by the Norwegian Research Council, project nr. 160130/V30.

One attempt to categorize image processing frameworks may be the following:

- Some address issues like software reusability through emphasis on image processing algorithm generics (templates). These contain independent building blocks, and the user can restrict the use of blocks to only the ones he needs. An example of this kind is the *Vigra Computer Vision Library* [5].
- Some libraries, like Java’s image processing library [6], attempt to be as general as possible. The user has access to rich functionality, even if he may only be interested in a small part of it.

RIM stands somewhere between these types. It does not attempt to be a set of loosely coupled general purpose algorithms, although parts of it may be extracted as a template library. It does not attempt to be a fully featured image processing framework either. It is a small set of high-level interfaces targeting component-oriented usage. The interfaces offer general image operations, particularly transcoding between widely used formats. These operations are abstract to the user in the form of an *Image Algebra*, a set of functions to compute new images as a function of other images. The image operations are polymorphic with respect to the concrete image formats.

2 The RIM Core

The core of the RIM framework is implemented in C++, see public header file [7]. It is at an experimental stage, so that not all parts of it have been optimised or tested. RIM does not link with other image processing libraries, and support for some image standards have been implemented from scratch. This was done in order to support optimisations for lazy evaluation and *runlength-based image processing*.

The interface to the RIM framework is inspired by Microsoft’s Component Object Model, or COM [8], in order to be programming language independent and target distributed applications. COM provides a standardized API in the form of the `IUnknown` interface, and all RIM functionality is based on COM interfaces offering this interface. Although the interfaces were implemented in C++, the COM interop system of .NET [9] can use these interfaces with its garbage collector to achieve a smooth integration with languages based on CLI (Common Language Interface), such as C#. C++ and Java are the languages currently supported by RIM. The Java interfaces are given in [10]. Interface naming conventions in this paper follow those in [10], with the exception that a class prefix is dropped.

One advantage with an interface-based API is that one can hide implementation strategies. Different image formats can for instance utilize data representations in different *domains* during image processing, the details of the domains and when they are chosen being completely hidden to the application developer. The most widely used image processing frameworks assume that a *raster representation* is used. RIM takes this further by using both runlength-based and raster-based internal representations [1], the choice depending on the image format. Some image formats and operations may be most efficiently processed when

a runlength-based internal representation is used, and [1] exploits this in terms of image transcoding. It was shown that more efficient processing is obtained when the input and output formats can efficiently convert between runlength representations and compressed data. GIF and bi-level TIFF were used as examples for such formats. One can utilize other internal representations also. Operating directly in the wavelet domain is for instance known to be more efficient for certain operations [11].

The high-level abstractions of RIM makes it suitable for use as a dynamic image library in a web application setting. RIM has been integrated with an *XML interpreter*, where different XML elements correspond to different RIM interface methods. Example XML files can be found in [12]. The XML interpreter has been integrated with an *Image Server component* [1] for prolonging the life span of ephemeral images likely to be used in the future. The Image Server is designed to host image requests for a map server, so it can be seen as an analogy to WMS. It is reviewed in section 4.1.

2.1 RIM Main Interfaces

Certain interfaces are of particular importance in RIM. The most fundamental interface is `Image`, which is the interface abstraction of an image format's read-only view to the image data. The `Image` interface contains methods for retrieving common image characteristics, like dimensions. An `Image` can offer other interfaces also, reflecting different aspects of the underlying image data. It may for instance be that the underlying image data is actually vector data. The `VectorSource` interface is then also offered. This offers vector-based methods, like functionality for processing objects like text, circles and lines. The `ColoredImage` interface is offered if a colour image is used. Interface inheritance relationships are summed up in figure 1.

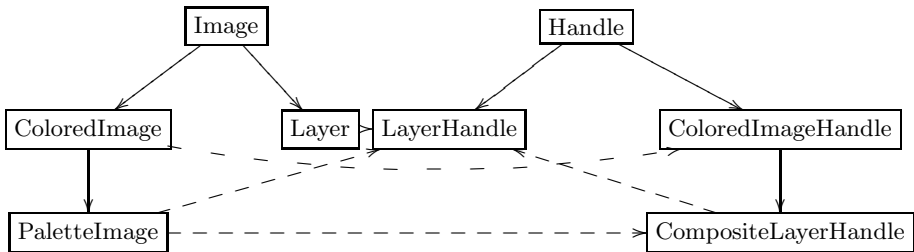


Fig. 1. `Image` and `Handle` hierarchies. Solid line represents inheritance. Dashed line represents ways of producing objects of the given types.

The `Image` interface provides a method for producing *references* to its image data. The references represent the context of image traversal, and are used when iterating a (potentially) compressed image. Many references could be created. Each can traverse the image data independently, thereby supporting concurrent image processing. The references offer the interface `Handle` (figure 1),

which supports functionality like rendering image data to an output buffer. In Java, there is a similar duality between the classes `Graphics` and `Image`. In the RIM implementation, classes implement these interfaces on a per-format basis. The inheritance hierarchy in figure 1, together with internal processing domains shared by many image formats, offer possibilities for code reuse. This is reflected in the relatively small code footprint achieved by RIM: The entire RIM dll is only about 450kb when compiled on win32 platforms.

`Handle` objects need not arise as references to image data. They can also arise from `Font` objects, which represent textual data. `Handle` objects can also be constructed from Image Algebra operations.

2.2 Image Algebra

Map images typically consist of a number of bi-level layers placed together. RIM supports bi-level images in the following way: If an `Image` originates from a bi-level file, it will offer the `Layer` interface. `Layer` objects support *boolean* operations. A particularly important boolean operation is *image difference*. Boolean operations are part of an important category, called *Image Algebra operations*. Image Algebra operations produce `Handle` objects from existing `Handle` objects. Other examples of Image Algebra operations supported in RIM are

- scaling, which produces a scaled `Handle` object,
- rotation,
- clipping,
- duplication,
- combining a set of `Handle` objects in a given Z-order,
- separating a colour-indexed image into `LayerHandle` objects,
- inversion (switching foreground and background in a bi-level image).

Image Algebra operations can be combined recursively to form a tree, for instance by taking image difference of scaled or rotated `Handles`. In such a tree, leaf nodes would correspond to what one may call *atomic Handles*. These include `Handles` which are references to image data. An Image Algebra tree using some of the listed operations is shown in figure 2. Note that Image Algebra operands can refer to either image data or vector data, opening up for applications to *hybrid formats* like SVG [13].

The common factor for Image Algebra operations is that new `Handle` objects are created. How this is done is up to the implementation, but it is recommended done without creating new image data segments. RIM is implemented with this in mind, for instance by performing operations like scaling and image differences with only small parts of the image loaded at any time.

If an intermediate Image Algebra result is reused more than once, it may be desirable to precalculate the Image Algebra to avoid performing repeated Image Algebra. A method in the `Handle` interface offers this functionality, and creates a compressed in-memory representation of the Image Algebra tree. The format used for this representation is at the discretion of RIM, and different formats are used for different image content: TIFF G4 is used for bi-level images, a

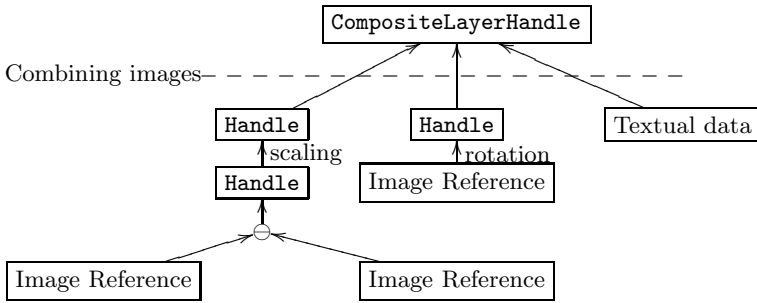


Fig. 2. A typical Image Algebra tree. \ominus represents image difference. Text data is placed on top of the image layers.

proprietary format is used for vector data. JPEG2000 is a natural candidate for colour images.

The functionality for compressing to an in-memory representation, along with the other Image Algebra operations, constitute a rather complete set of image operations. Performing Image Algebra raises a string of performance issues, like how Image Algebra trees can be transformed into equivalent trees more suitable for processing. RIM implements several such optimisations.

2.3 I/O Support in the RIM Framework

RIM supports GIF, BMP and TIFF input. TIFF input is analysed in [1], where the focus is on TIFF G4 [14]. An API method exists which creates an **Image** object from file name and file type identifiers. Depending on the image type, this object may offer any of the interfaces already discussed.

RIM supports GIF, TIFF, lossless JPEG2000, JPEG and PNG output. The PNG implementation is based on the libpng reference library [15]. The RIM framework supports different types of output through a method taking an output format identifier as parameter. This method creates an object offering the **Renderer** interface. The **Renderer** interface has a method which, for a selected image region, incrementally renders compressed output to a buffer. The method signature is similar to the **read**-methods of `java.io.InputStream` classes in Java: A parameter indicates the size of the buffer to read from, and another parameter indicates the number of bytes actually read. Such a method signature frees us from the underlying file system: The output buffer can for instance be drained onto a network connection, enabling integration with web servers. Another advantage is that one is offered natural support for splitting output in logical units since the method can produce output in parts. Logical units for different image formats could be blocks (used by GIF), chunks (PNG) or packets (JPEG2000). Java also uses **InputStreams** for image processing purposes, for instance for the *deflate compression algorithm*.

Prior to rendering compressed output, one must restrict compression to a concrete region, and the **Handles** to render must be added. **Handles** which are

results of Image Algebra expressions are typically added, and the order they are added dictates the Z-order. A typical application can have a colour image or a set of bi-level images as background, and have text fragments or small bitmap images anchored at designated positions. Bitmaps may be used to represent some kind of user interaction (like zoom or pan), so this could constitute a user interface. Example XML is listed below:

```
<?xml version="1.0" encoding="UTF-8" ?>
<visalg>
  <coloredsection color="beffe9">
    <file x0="0" y0="0" laysf="1" name="l1.tif" format="3"/>
  </coloredsection>
  <coloredsection color="ffd1bf">
    <file x0="0" y0="0" laysf="1" name="l11.tif" format="3"/>
  </coloredsection>
  <coloredsection color="000000" static="true">
    <text height="16" width="8" text="Test" x0="10" y0="70"/>
  </coloredsection>
  <coloredsection color="00ff00" static="true">
    <file x0="10" y0="40" laysf="1" name="rimtool.bmp" format="2"/>
  </coloredsection>
</visalg>
```

When RIM's XML interpreter processes this, two TIFF layers overlaid with a black text segment and a green bitmap will be produced (figure 3).

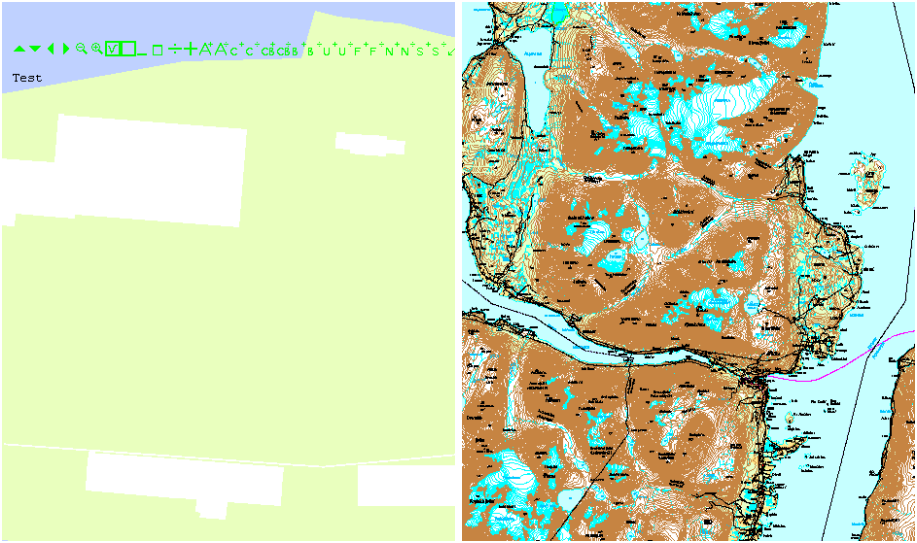
3 Applications of the RIM Framework

A useful and simple application of RIM is layer separation. One of the dashed arrows in figure 1 represents layer separation, so that occurrences of a single colour in a colour-indexed image may be obtained as a dedicated object. This object can be compressed to an in-memory representation, which may be desirable to avoid repeated colour separation.

Performance results are here obtained for different image formats using RIM. The image formats which will be used are GIF, PNG and lossless JPEG2000. GIF and PNG are perhaps the most widely used formats for exchange of losslessly compressed images on the world wide web, while JPEG2000 is the emerging standard for both lossy and lossless compression. Measurements use the same test images as in [1], i.e. two images of different parts of Norway comprising of 19 TIFF G4 bi-level layers. One of these is 7469×8886 pixels in size (figure 3). The test images have tile dimensions of 512×512 , and tests are performed on the tiles separately to obtain a high number of tests. XML files written for the tests are listed in [12].

3.1 Comparison of Performance for Different Output Formats

Performance in terms of clock cycles should be higher when little detail is present in the image. For RIM, this is verified in the first plot in figure 4, where *accumu-*



(a) Output for the XML example listed (b) Layered image of Lyngen, one of the test images used in this paper

Fig. 3. Images used in this paper

lated runs per line is plotted against clock cycles. Accumulated runs per line [1] measures the level of image detail in the form of counting the number of runs per line for all layers. The connection between performance and image detail is best seen for GIF and PNG. GIF comes out best in terms of performance, as it has the least complex algorithm. For JPEG2000, two main components have impact: The embedded block coder (EBCOT), and the Discrete Wavelet Transform (DWT). The DWT has not been applied in the plot, so the poor performance of JPEG2000 as compared to GIF has to do with the complexity of the embedded block coder. The most expensive part of a PNG compressor is the *matching algorithm* part of *deflate*. If much time is spent matching previous combinations of pixels, compression is improved. The PNG compressor used here is more concerned about compression than performance, which is reflected in poor performance numbers when compared to GIF.

It may be that compression of bi-level images is of interest. According to [16] chapter 16.3, JPEG2000 outperforms GIF when it comes to compression at low bit-depths, and is comparable to JPEG-LS and TIFF G4 (for bi-level images). The second plot in figure 4, generated by using just one layer (rich in content) in the test image, supports this statement.

3.2 JPEG2000 Compression Strategies

JPEG2000 is flexible when it comes to techniques which can improve compression. Palette mode can be used for images with a limited number of colours. Palette-based JPEG2000 can improve compression considerably for two reasons:

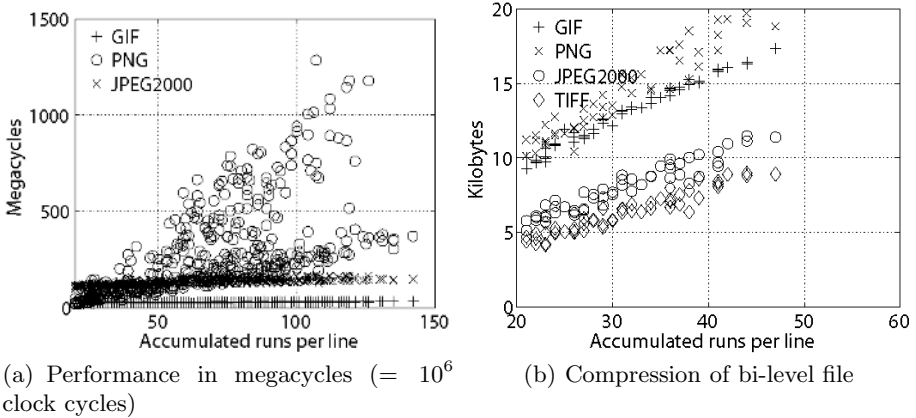


Fig. 4. Comparison of widely used image formats using RIM

First of all, bit-depth and the number of components are reduced. Secondly, palette-indices can be reorganized. This can be exploited by the JPEG2000 compression algorithm, since the JPEG2000 block coder is bit-plane oriented and gives higher compression in areas with low bit-plane complexity. Reorganizing palette indices for some image formats has been exploited in [17]. The figures in this paper have used a simple palette reorganization, in which the background is assigned palette index 0, and the next colours are assigned indices in alternating and increasing order around 0. Both PNG and JPEG2000 support palette mode, and so does RIM for both these formats. Comparison with and without palette mode is done in figure 5 for these two file formats.

Both JBIG and JPEG2000 can apply multi-resolution transforms. [16] notes that the reversible wavelet transform JPEG2000 uses is primarily designed for continuous-tone imagery. One would therefore expect that compression would

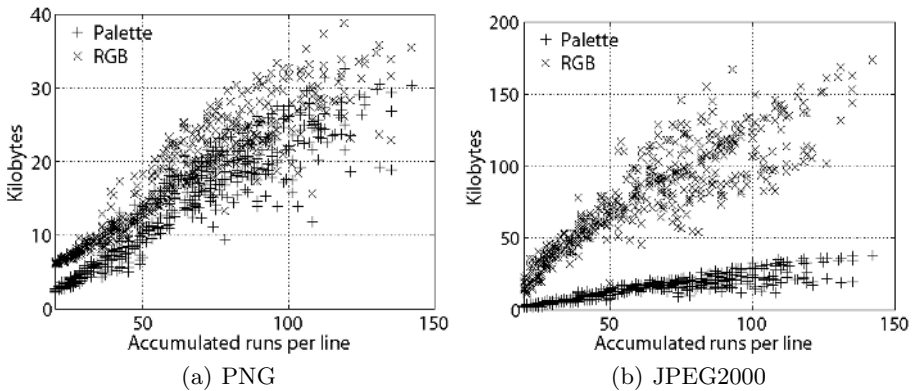


Fig. 5. Comparison of compressed file sizes for palette-based and RGB-based compression. PNG and JPEG2000 are used.

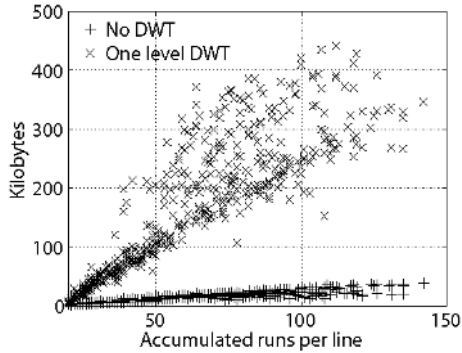


Fig. 6. Comparison of compressed file sizes for JPEG2000 with no DWT and one level DWT

suffer somewhat for our type of images when different resolutions are used. This is verified in figure 6, where compressed file sizes for zero and one DWT levels are compared. RIM uses a `Config` interface for image format specific configuration. For JPEG2000, this supports setting tile sizes, block sizes, *progression order* [16] and the number of DWT levels. The JPEG2000 `Config` interface is here used to set the number of DWT levels.

4 Integration of RIM with Other Component Libraries

RIM can easily be integrated with components like web servers and GUI libraries. Qt [18] is a C++ class library for writing GUI applications. It has been used to build the popular open source KDE desktop environment for Unix. Making a *scrollable component* with QT boils down to subclassing the class `QScrollView`, and implementing the method `drawContents` to draw the image contents of the current part of the image. An example file in [19] sketches how this can be done using RIM. The RIM framework can also be integrated with Java Swing components or Java servlets in a similar way. An example file in [19] sketching this is also listed.

4.1 Integration with the Image Server

The Image Server acts as a cache for frequently accessed files, and as a front-end to RIM. A typical use of the Image Server is to extract a small part of a large image on request from a web server. The Image Server ensures that frequently requested parts are readily available in shared memory. The architecture used by the image server is shown in figure 7.

The Image Server could also be used as a cache for compressed representations of the most commonly used Image Algebra requests. Another possible use could be to serve as information holder for occurrences of colours within different parts of the image. Such information can be used in the process of improving compression.

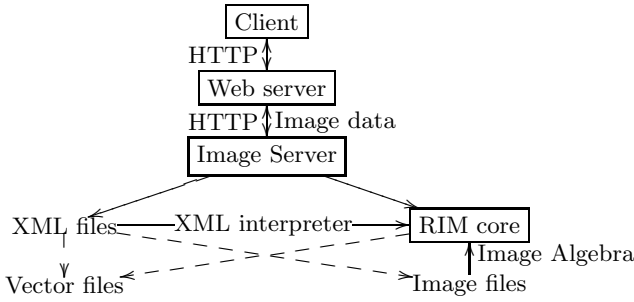


Fig. 7. The Image Server architecture

JPIP [20] is one of the more recent extensions to JPEG2000. It defines a protocol for scalable delivery of JPEG2000 data in client-server systems. Supporting the JPIP protocol is another possible use of the Image Server.

5 Other Work

The Image Server generates images from image description files using XML. Other dynamic image libraries have also been developed for use in web development settings similarly to RIM. An example is the *gd* library [21], which has been integrated with the *fly* *gd* command interpreter. Separate *gd* commands exist for different drawing primitives, so that image processing can be embedded in scripting languages. This is similar to the way XML is used by the Image Server. The RIM API supports the most common drawing primitives, like circles, lines and text, so the RIM XML interpreter supports similar functionality to *gd*.

6 Conclusion

A small dynamic image processing library has been demonstrated. It was argued that the library meets low-memory demands imposed in dynamic image processing. It was also shown how the library can be used to demonstrate properties about widely used image standards, and easily be integrated with other GUI component libraries. It was also demonstrated that RIM can handle different image formats in a completely transparent manner, and how RIM's support for Image Algebra makes it a very general tool.

Results in this paper were obtained with an Intel Pentium M processor with 1600MHz clock speed, L2 cache size of 1MB and 512 MB RAM. All tests were run under Windows XP, and all programs were compiled with Microsoft Visual C++.NET 7.1.

Acknowledgement

I give my sincere thanks to Stein Jørgen Ryan for helpful discussions on different topics presented in this paper.

The work in this paper is partially based on the RIM library from Raster Imaging AS (www.rasterimaging.com) which provides high performance imaging technologies. The post.doc project carried out by Dr. Øyvind Ryan at the University of Oslo has enhanced this implementation, and added algorithms for improved performance and scalability with regards to server applications and memory consumption.

References

1. Ryan, Ø.: Efficient implementations of operations on runlength-represented images. Submitted to the 14th European Signal Processing Conference, Eusipco 2006 (2006)
2. Denning, P.J.: The working set model for program behavior. *Communications of the ACM* **11** (1968) 323–333
3. Denning, P.J., Schwartz, S.C.: Properties of the working-set model. *Communications of the ACM* **15** (1972) 191–198
4. Open Geospatial Consortium Inc.: WMS specification. <http://www.opengis.org>. (2006)
5. Köthe, U.: The Vigna computer vision library. kogs-www.informatik.uni-hamburg.de/~koethe/vigra/. (2005)
6. Sun Microsystems: Java Image I/O API. java.sun.com/j2se/1.4.2/docs/guide/imageio/. (2002)
7. Raster Imaging AS: RIM framework C++ header file. www.ifi.uio.no/~oyvindry/rim/rim.h. (2006)
8. Microsoft: COM. www.microsoft.com/com/default.msp. (2006)
9. Löwy, J.: *Programming .NET Components*, 2nd Edition. O'Reilly Media (2005)
10. Raster Imaging: javadoc for RIM. www.ifi.uio.no/~oyvindry/rim/javadoc/. (2006)
11. Drori, I., Lischinski, D.: Fast multiresolution image operations in the wavelet domain. *IEEE Transactions on Visualization and Computer Graphics*. **9** (2003) 395–412
12. Raster Imaging AS: Example xml files. www.ifi.uio.no/~oyvindry/rim/. (2006)
13. W3C Consortium: SVG specification. www.w3.org/Graphics/SVG/. (2006)
14. CCITT: Recommendation T.6. Facsimile Coding Schemes and Coding Control Functions for Group 4 Facsimile Apparatus. (1985)
15. libpng.org: libpng, reference library for reading and writing PNG. www.libpng.org. (2001)
16. Taubman, D.S., Marcellin, M.W.: *JPEG2000. Image compression. Fundamentals, standards and practice*. Kluwer Academic Publishers (2002)
17. W. Seng, J.L., Lei, S.: An efficient color re-indexing scheme for palette-based compression. *Proc. IEEE Int. Conf. Image Proc.* **3** (2000) 476–479
18. Dalheimer, M.: *Programming with Qt* (2nd Edition). O'Reilly Media (2002)
19. Raster Imaging AS: RIM framework example files. www.ifi.uio.no/~oyvindry/rim/. (2006)
20. The JPEG Committee: ISO/IEC 15444-9:2005, Information technology - JPEG 2000 image coding system: Interactivity tools, APIs and protocols. (2005)
21. boutell.com: The GD graphics library. www.boutell.com/gd/. (2006)

A Proposal Method for Corner Detection with an Orthogonal Three-Direction Chain Code

Hermilo Sánchez-Cruz

Centro de Ciencias Básicas. Universidad Autónoma de Aguascalientes
Av. Universidad 940, Col. Universidad, CP. 20100
Aguascalientes, Aguascalientes. México. Fax: (52 449) 9 10 84 01
`hsanchez@correo.uaa.mx`

Abstract. Only three set of pattern chain elements to detect corners in irregular shapes are introduced. A code based on three orthogonal change directions, when visiting a contour shape, are used. Previous approaches for detecting corners employ eight different symbols and usually compute angles and maximum curvature. The three basic pattern contour chain elements, founded in this paper, represent changes of direction in the contour curves, requiring few computing power to obtain corners. Also, we have found that the method is independent of shape orientation.

Keywords: Shape corner; Contour; Chain element; Freeman chain code; Symbol chain code; Pattern substrings.

1 Introduction

Nowadays, corner detection of shape objects is an active field in object recognition and image retrieval. In literature, usually the aim in obtaining corner points by computing angles of curvature on the contours of shapes is studied and representing discrete contours by Freeman chain codes. Freeman and Davis [1] proposed to find corners by computing incremental curvature to represent contour shapes by an eight-direction chain code. Since then, many authors have suggested to use this code when representing contour shapes. Part of the algorithm presented by Teh and Chin [2] consists on computing the curvature of contour points and detecting corners by a process of nonmaxima suppression. Liu and Srinath [3] have compared a number of corner detectors due to Medioni and Yasumoto [4], Beus and Tiu [5], Rosenfeld and Johnston [6], Rosenfeld and Weska [7] and Cheng and Hsu [8]. All those authors represented samples of shapes through a sequence of eight direction changes from 0-7, known as the Freeman Chain Code [9].

Wu [10] proposed an adaptive method to find local maximum curvatures of digital curves. Sobaina and Evans [11] described a corner detection from segmented areas using mathematical morphology employing paired triangular structuring elements.

Basak and Mahata [12] developed a connectionist model along with its state dynamics for detecting corners in binary and gray level images. We have studied that it is suitable to represent any binary closed shape with binary resolution cells, and using only three symbols of a chain code, without loss of information

[13]. This method of chain code is sufficient to represent binary shapes and represents low cost in storage memory. Techniques due to Freeman chain codes in finding corner detection are based in eight different directions (see Fig. 1a).

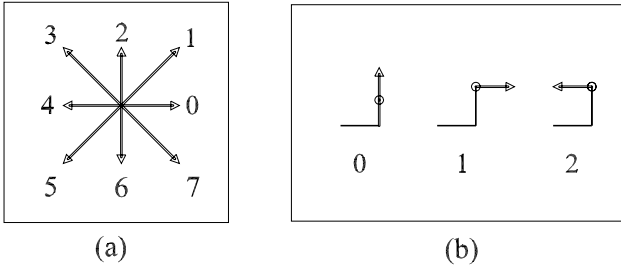


Fig. 1. Two different chain codes: (a) The eight different directions given by Freeman chain code. (b) The three orthogonal change directions.

We propose here to use a method of only three relative direction changes given by Fig. 1b. An advantage in using three symbols is its low storage power, as can be seen by the recent work duo to Sánchez-Cruz & Rodríguez-Dagnino [13]. They found that coding with three symbols is sufficient to represent binary shapes saving storage efficiently. Recently Yong Kui Liu & Boruk Zalik[15], found efficient storage properties by using Huffman coding applied on change directions of Freeman chain code.

For each orthogonal change direction code, chain segments are divided in three parts (given in Fig. 1b): a *reference segment* (in Fig. 1b appears as horizontal segment in each code), a *basis segment* (perpendicular to reference segment) and a segment indicating a direction change with regard to reference segment.

The meaning of the three symbols (see Ref[14] for 3D case), given by the set $\mathcal{C} = \{0,1,2\}$, is as follows: the element 0 represents the direction change which means to “go straight” through the contiguous straight line segments following the direction of the last segment; the ‘1’ indicates a direction change upward with regard to the reference segment; and ‘2’ means to “go back” with regard to the sense of the reference segment. In this work we have noticed that when the symbol ‘2’ appears in a contour shape, can easily indicate an existing corner. In Section 2 definitions concerning to this article are presented, seeking the problem as a pattern substring search. In Section 3 some rules to detect shape corners are proposed; in Section 4 experimental proving of postulated rules are applied on some binary shapes; in Section 5 rotation independence is analyzed, and in Section 6 we give some conclusions.

2 Some Definitions

Our proposal method considers to find a specific set of pattern substrings of length l , trying to find all those substrings in a shape contour coded by a chain

that match with those patterns. Let us consider, for example $l = 11$ as the length of the substrings. Which substrings are all composed of 11 symbols and which of them are considered corner chains? In fact, there are substrings composed of 11 symbols, of course, not all are considered corner chains due to its low curvature or because the region they are associated in the contour shape is not “well behaved”, as we explain at once.

Let \mathcal{P} denote the *complete chain code* (or simply *chain code*) associated to the shape contour, given by the string of symbols p_i of eq(1).

$$\mathcal{P} = p_1 p_2 \cdots p_n, \tag{1}$$

and P the contour discrete perimeter, given by the number of symbols of the chain code.

Consider a substring template of l symbols: $C \in \mathcal{P}$, given by eq(2).

$$C = a_1 a_2 \cdots a_l, \quad l \ll P, \tag{2}$$

as a *contour chain element*, or simply: *chain element*, this is, a small piece of contour from the whole shape contour.

Let us consider $m = l/2$ the middle point of a substring of size l , so that a_m , the pivot, be the center of the substring. It is possible to associate a pair of line segments to any chain element. They can be drawn up from to the opposite end points, producing an angle φ . We define a *well behaved* chain element when an angle has been subtended by a pair of associated line segments such that the chain does not form loops. Observe a sample of chain elements and their corresponding visual meaning in Fig. 2.

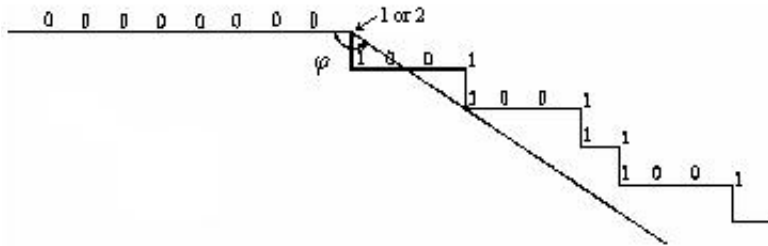


Fig. 2. Angle φ of curvature. Consider a 5-neighborhood of pivot (1+2): 00000(1+2)10011.

We define a *neighborhood of radii r* , when considering a piece of the complete string; this region is composed of a small number of symbols in comparing with the whole contour chain code, r symbols on one side of a particular pivot symbol, and r symbols on the other side of the pivot symbol. Fig. 2 presents an example of neighborhood of radii 5: chain elements having 00000 in its left first part, 10011 in the right part, and 1 or 2 (1+2 to abbreviate) as a pivot symbol could appear on the contour shape.

We propose to save calculation of corner-angles or curvature changes directly. Instead we give a family of substrings that represents high curvature. Well behaved substrings should not be considered a corner chain when their corner chains associated angles are so much obtuse.

Another definition we need is a *well behaved contour shape*, this is, a contour shape having been smoothed in such a manner that there is no noise or local defects.

3 Rules for Detecting Chain Corners

In this work we did our experiments with a vicinity of eleven symbols in chain elements giving good results in finding chain corners.

A way to obtain a complete set of templates considered chains corners, is to search all the substrings arrays composed of l symbols from the set $\mathcal{C} = \{0,1,2\}$, calculate the angle associated to each substring and apply the threshold to see if it is a chain corner. But we propose a small enough set of template substrings to find the evident chain corners from an arbitrary set of 2D shapes. To find a group of pattern substrings or pattern chain elements considered as chain corners is to focus in a vicinity of each change code contour, by for example eleven chain segments, nine of them labeled with symbols, representing orthogonal direction changes. The first two are called reference segment and basis segment, respectively. There are a huge number of combinations given by nine symbols (11 segments) and we are interested on finding chain elements that have no loops. Even more, fixing the reference segment of the chain, there are 3^9 combinations duo to the other nine chain directions. Fig. 3 presents part of these combinations.

By analyzing the different chain sets mentioned, we have observed that pattern substrings representing corners. Parameters we have to take into account are the next:

l : states for the size of substring.

q : represents “many” times a symbol is repeated in a substring. This quantity depends on resolution of binary object. By *many* we define that the number of symbols is greater than $l/4$, so $q \in (l/4, l]$.

Part of the study made to find a simple pattern of substrings that represent corners, was to find that we have to consider only the cases when there is an appreciable direction change when visiting the discrete contour. At first glance, this happens with high probability when a symbol ‘2’ from the orthogonal changes appears. As was introduced in Section 1, this symbol represents “go back” when covering the shape contour, indicating a corner shape. Of course, in a given situations, where there is some noise or local shape defects, ‘2’ symbol should not constitute a perceptible corner. So, our first pattern chain class to be proven is composed of a string of 1s or 0s, then a ‘2’ symbol and then a sequence of 1s or 0s again.

Another pattern chain elements, that represents contour change directions, occurs when there are many 0s (with possibly some pairs of 1s) following a ‘1’

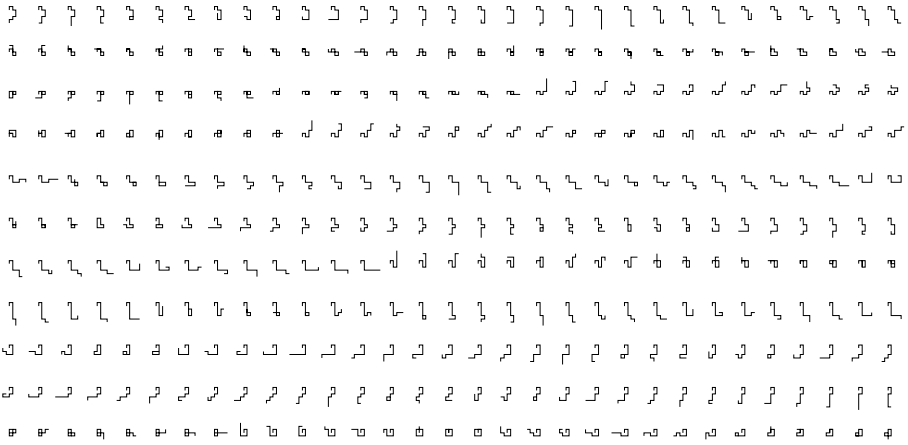


Fig. 3. Part of the complete set of 3⁹ substrings

as a middle symbol of the chain element, following again many 0s (with possibly some pairs of 1s).

The other pattern chain elements represents only changes of directions of the contour shape. Changes of directions occur simply when there are a substring of many 1s, with possibly some 0s, following a substring of many 0s (with possibly some 1s); or viceversa, a substring of many 0s (with possibly some 1s) following a substring of many 1s (with possibly some 0s).

So, to simplify the pattern of chain elements that correspond to chain corners, we are talking about pattern strings of discrete chain corners, postulated by next regular expressions:

$$\begin{aligned}
 S_1 &= (0 + 1)^{l/2}(\mathbf{2})(0 + 1)^{l/2}, \\
 S_2 &= (0^q + 1_p)^{l/2}(\mathbf{1})(0^q + 1_p)^{l/2}, \\
 S_3 &= (0^q + 1_p)^{l/2}(\mathbf{1})(0 + 1_p^q)^{l/2} + (1^q + 0)^{l/2}(\mathbf{1})(1 + 0^q)^{l/2},
 \end{aligned}
 \tag{3}$$

where q represents *many* symbols, p states for a pair of symbols: $p = \{0, 2\}$. Rule S_1 states that when a symbol 2 appears, the chain element can be considered a corner chain. Rule S_2 means there are many zeros in both sides of the middle point of the the chain element. Rule S_3 means that substring has many zeros or ones in the first part of the chain element and many ones or zeros in the second part of the chain. See Fig. 4 for a sample of these rules. Our proposed method relies on looking for these pattern substrings on any contour shape.

We consider shapes represented by resolution cells, each having a value 0 or 1. For the implementation of an algorithm to encode this shape we have to visit the ones that represent the contour shape, i.e., the ones of the boundary. We follow the contour of the shape clockwise sense, updating in every step, the reference segment with the contiguous segment, and giving one of the three symbols according to each orthogonal change direction (see Ref[13] for a detailed explanation).

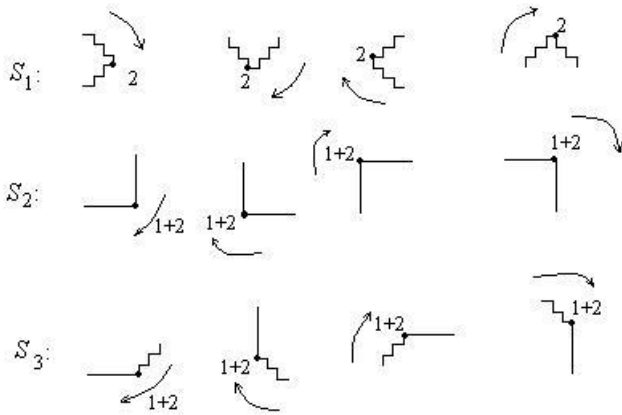


Fig. 4. Samples of the three types of corners, each invariant under rotations transforms

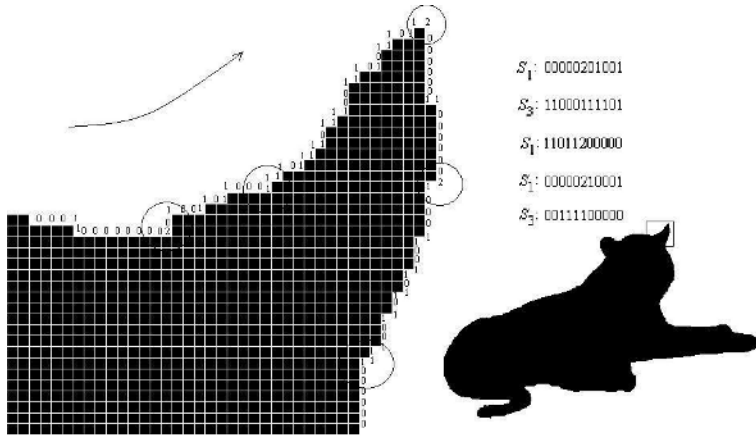


Fig. 5. Examples of chain elements, covering the contour on clockwise sense. The grid is part of the inner shape.

The object shape is confined to a minimum rectangle that is visited line by line, from left to right and from top to bottom. The first cell resolution, of the object to be visited, is that which appears at the leftmost and highest part of the occupied region. Fig. 5 shows part of a contour shape and examples of chain elements coded by the three symbols of the orthogonal directions given in Fig. 1b. Given this representation, we can reconstruct the original image by interpreting the code of every symbol in terms of the direction changes that can follow.

Finally, the pattern substrings, S_1 , S_2 and S_3 are parsing the resulting chain string of the complete contour.

4 Experiments

Consider the set of four shapes $S = \{Irregular\ shape, Circles, Hammer, Tigger\}$ showed in Fig. 6 as binary objects.

Consider *Irregular shape* object and its corresponding chain code (Fig. 7), 51 chain corners were found in its conotur shape. Some of them are so closed, in

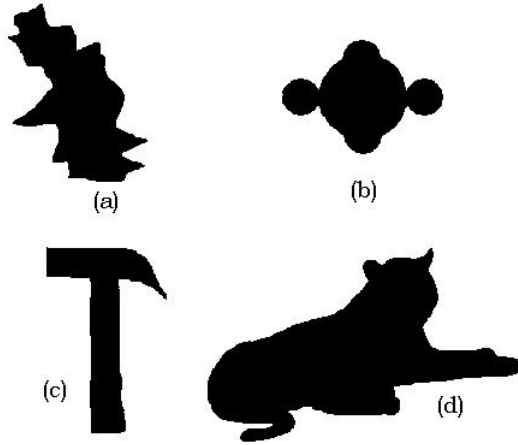
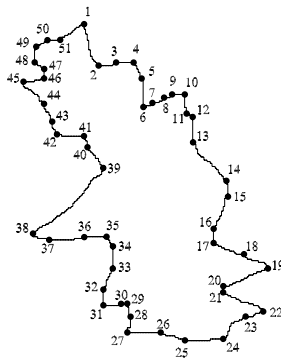


Fig. 6. Four shapes: (a) *Irregular shape*; (b) shape of intersecting *Circles*; (c) *Hammer*; (d) *Tigger*



(a)

```

011112001100110001100000110001101111111110000000021011000000000211110110110001111000000000000
000102100111100011011110001111000000200000001100010111110000000000000111100110011111010101
1011101011110111110111101110110000000211100001100110011011110011011000000002101101101100111100
1100011111100110001100011111020110110011011001110011100110110011011000100021101101100110111000011
0011001111010201100111100001101101100110101100110011001110110000000110000000000000210001101100001
100110000000000000000000000001002010000002100000110100021000000000020000000000210110001101101100110
00000000000021011011110000000000002100000110000000000000021000000110010211111111011011011011
111011110111110111011111111111111111111111111111111111111111111111111111111111111111111111111
000000001100111011110001100011010101010101010101011111111111111111111111111111111111111111111
1100000002101110110110110110000000111101101101111111
    
```

(b)

Fig. 7. *Irregular shape* and its corresponding 3-symbol chain code: (a) the shape; (b) its chain code

Table 1. Chain elements encountered from the Fig. 7 that belong to one of the two classes of chain patterns postulated. Corners 5,9,27,41 and 46 are split by two closed chain corners.

<i>Num corner</i>	<i>Chain element</i>	<i>Class pattern</i>	<i>Num corner</i>	<i>Chain element</i>	<i>Class pattern</i>
1	01111200110	S_1	12	01111100000	S_3
2	11111100000	S_3	13	00000111100	S_3
3	00000210110	S_1	14	11101100000	S_3
4	00000211110	S_1	15	00000211100	S_1
5a	11000111110	S_3	16	01101100000	S_3
5b	00111100000	S_3	17	00000211011	S_1
6	00010210011	S_1	18	10000111111	S_3
7	01111100011	S_3	19	11110201101	S_1
8	10111100011	S_1	20	11000100021	S_2
9a	11000111100	S_3	21	01000211011	S_1
9b	00111100000	S_3	22	11010201100	S_1
10	00000200000	S_1	23	10000110110	S_3
11	11000101111	S_3	24	11101100000	S_3
12	01111100000	S_3	25	00000210001	S_1
26	01101100000	S_3	39	10110210110	S_1
27a	00000100201	S_2	40	11011000111	S_3
27b	00100201000	S_1	41a	11000111110	S_3
28	00000210000	S_1	41b	11110100000	S_3
29	00000110100	S_2	42	00000110011	S_3
30	01000210000	S_1	43	10111100011	S_3
31	00000200000	S_1	44	11000110101	S_3
32	00000210110	S_1	45	11100200000	S_1
33	11001100000	S_3	46a	11000110100	S_2
34	00000210110	S_1	46b	00110100000	S_2
35	11111100000	S_3	47	00000211110	S_1
36	00000210000	S_1	48	11111100000	S_3
37	00000210000	S_1	49	00000210111	S_1
38	10010211111	S_1	50	01101100000	S_3
			51	00000111101	S_3

such a manner that their corresponding pivots are in the neighborhood of each other, in this case we could define only one corner. In Table 1 is listed each of the chain elements and the corresponding class pattern given by eq(3) of the *Irregular shape* contour.

From Figures 7 to 10, contour shapes of the set \mathcal{S} and their corresponding chain codes are presented. They also show the results of applying the method proposed to search chain corners given a substring length of $l = 11$.

6 Conclusions

To save time and memory storage to manage this kind of objects, we used three symbols that represent orthogonal directions when covering contours of binary shapes. With this method we have found shape corners including there where is apparently circular, like happening with figure constructed by intersecting circles. We found three classes of patter substrings to obtain the most important shape corners in contour shapes, preventing to compute angles and curvatures directly; so we have presented a new research topic. Also, we have analyzed that the proposed metod is invariant under rotations of shape contours. As future work most be studied if this method is invariant under scale transforms. A universal and simplified set of pattern substrings, comparing with other chain codes in literature is suggested to be investigated.

Acknowledgments

We would like to thank PROMEP program and CONACyT council for their support in finishing this work.

References

1. H. Freeman and L. S. Davis, A Corner-Finding Algorithm for Chain-Coded Curves. *IEEE Trans. Comput.* 26: (1977) 297-303.
2. C-H. Teh, and R.T. Chin, On the Detection of Dominant Points on Digital Curves. *IEEE Trans of Pattern Anal and Mach Int.* 11 (8) (1989) 859-872.
3. Hong-Chih Liu; M.D. Srinath. Corner Detection From Chain-code. *Pattern Recognition.* 23 (1/2) (1990) 51-68.
4. G. Medioni; Y. Yasumoto. Corner detection and curve representation using cubic B-Splines. *Comput. Vision Graphics Image Process.* 39: (1987) 267-278.
5. H.L. Beus; S.S. H. Tiu. An improved corner detection algorithm based on chain-coded plane curves. *Pattern Recognition.* 20 (1987) 291-296.
6. A. Rosenfeld; E. Johnston. Angle detection on digital curves. *IEEE Trans Comput.* 22: (1973) 875-878.
7. A. Rosenfeld; J.S. Weszka. An improved method of angle detection on digital curves. *IEEE Trans. Comput.* 24: (1975) 940-941.
8. F. Cheng; W. Hsu. Parallel algorithm for corner finding on digital curves. *Pattern Recognition Lett.* 8: (1988) 47-53.
9. H. Freeman. On the Encoding of Arbitrary Geometric Configurations, *IRE Trans. on Electr. Comp.* 10 (2) (1961) 260-268.
10. W. Wen-Yen. An adaptive method for detecting dominant points. *Pattern Recognition.* 36 (2003) 2231-2237.
11. A. Sobaina & J.P.O. Evans. Morphological corner detector using paired triangular structuring elements. *Pattern Recognition.* 38 (2005) 1087-1098.
12. J. Basak and D. Mahata. Connectionist Model for Corner Detection in Binary and Gray Images. *IEEE Trans. on Neural Net.* 11 (5) (2000) 1124-32.

13. H. Sánchez-Cruz; R. M. Rodríguez-Dagnino. Compressing bi-level images by means of a 3-bit chain code. *Optical Engineering*. SPIE. 44 (9) (2005) pp 1-8. 097004.
14. E. Bribiesca. A chain code for representing 3D curves. *Pattern Recognition*. 33(5)(2000),755-765.
15. Yong Kui Liu; Boruk Zalik. An efficient chain code with Huffman coding. *Pattern Recognition* 38 (4) (2005) 553-557.

A Charged Active Contour Based on Electrostatics

Ronghua Yang, Majid Mirmehdi, and Xianghua Xie

Department of Computer Science, University of Bristol, Bristol, BS8 1UB, UK
{ronghua, majid, xie}@cs.bris.ac.uk

Abstract. We propose a novel active contour model by incorporating particle based electrostatic interactions into the geometric active contour framework. The proposed active contour, embedded in level sets, propagates under the joint influence of a boundary attraction force and a boundary competition force. Unlike other contour models, the proposed vector field dynamically adapts by updating itself when a contour reaches a boundary. The model is then more invariant to initialisation and possesses better convergence abilities. Analytical and comparative results are presented on synthetic and real images.

1 Introduction

Ever since the introduction of the parametric snake [1], deformable models have received much attention for region segmentation and object detection. The geodesic active contour [2] is a significant improvement over the parametric snake in that it can naturally handle topological changes. However, it still suffers from drawbacks such as edge leakage and sensitivity to initialisation. There have been many efforts in improving both parametric and geometric snakes, for example by introducing region-based features to make the model more robust to initial conditions [3, 4, 5]. One significantly improved parametric model is the Gradient Vector Flow (GVF) snake [6] which uses a bi-directional external force field that provides long-range capture of object boundaries from either side. One of its main drawbacks however is that the contour does not propagate where the vector flows are tangent to the contour or diverge within a neighbourhood. One improvement of the geometric snake model is the GVF geodesic snake [7] which integrates the GVF with a geometric contour formulation and introduces an adaptive balloon force to help propagate the contour when the vector flows are tangent to the contour. This allows it to outperform the GVF snake while also benefitting from topological freedom. However, it is still unable to propagate through the points where the GVF field has large divergence which form in homogeneous areas depending on object topology. Therefore, the contours must be initialised with great care in order to avoid getting trapped at these points.

Recently, a new formulation for a “deformable model” based on charged particle dynamics, founded on electrostatics and particle movements, and called the Charged Particle Model (CPM), was introduced by Jalba et al. [8]. CPM can

capture object boundaries over the entire image with a set of free charged particles. These are attracted by object boundaries via an image-based force field, while at the same time being repelled from one another by a charged particle-based force which constantly imposes on the particles to advance them along object boundaries. While an initialization step is still required, it is certainly less pivotal than in the contour model. However, this particle model a) can not guarantee continuous and closed final contours, b) does not stabilise as there is no effective stopping term, and c) is computationally intensive.

In order to overcome the common drawbacks in the traditional deformable contour model and the deformable particle model, we propose a new framework by introducing particle based electrostatics into active contour propagation that incorporates the advantages of both contour and particle based models. We refer to this as CACE, a Charged Active Contour based on Electrostatics. CACE can detect object boundaries via contour propagation under the influence of a bi-directional force field that simulates the electrostatic interaction between an image-derived point charge field and a charged contour. In other words, the force consists of boundary attraction and competition terms that lead the contour towards object boundaries. CACE is much faster and more efficient in convergence than CPM. More importantly, it eliminates CPM's tendency to sometimes result in open contours. CACE also has significant advantages over the geodesic and GVF geodesic snakes in that it is more robust to initial placement and is able to handle objects of more complicated topology, e.g. those with narrow parts.

As electrostatics is the starting point of our work, we will review its key concepts in Section 2, along with a brief introduction to CPM and its shortcomings. In Section 3, the construction of our proposed model is discussed, with experimental results shown in Section 4. Section 5 concludes our work.

2 Background

An electrostatic field \mathbf{E} is defined as the electrostatic force upon a unit charge due to other charges. Suppose there is a distribution of N point charges c_1, c_2, \dots, c_N fixed at locations $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ respectively in a 2D space. According to Coulomb's Law, the electrostatic field at location \mathbf{x} in this 2D domain is given as:

$$\mathbf{E}(\mathbf{x}) = \sum_{i=1}^N \frac{c_i}{4\pi\epsilon_0} \frac{\mathbf{x} - \mathbf{r}_i}{|\mathbf{x} - \mathbf{r}_i|^3}, \quad \mathbf{x} \in \mathbf{X}, \quad (1)$$

where ϵ_0 is the permittivity of free space and \mathbf{X} is the set of all possible locations in this 2D domain. Thus \mathbf{E} is a vector field that has a force at every location in \mathbf{X} . If a test charge e is placed in the field at location \mathbf{x} , the electrostatic force put upon it can be obtained by:

$$\mathbf{F}(\mathbf{x}) = e\mathbf{E}(\mathbf{x}). \quad (2)$$

It is important to note that the electrostatic force acting on the test charge e is merely the superposition of separate electrostatic forces imposed by every

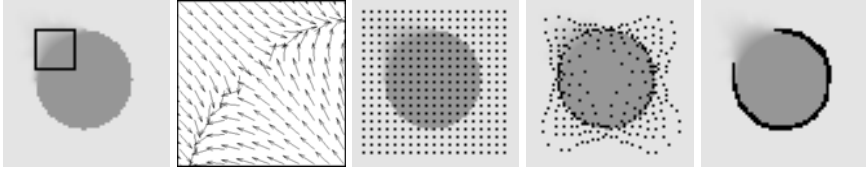


Fig. 1. Columnwise from left: synthetic circle image with highlighted weak edge area, the normalized Coulomb¹ force field in box area, initialized CPM, instance of particle movements, and the final CPM result

fixed charge c_i . This implies that the force of charge c_i upon test charge e is not influenced by the presence of any other fixed charges in space. This principle enables us to compute forces from different sources separately and control their contribution to our charged contour model.

CPM [8] is a particle model built on the simulation of particle movements in an electrostatic field. A set of positively charged free particles is placed in a field distributed with negative fixed charges proportional to the input image edge strength. As the particles have the same polarity to each other, and opposite polarity to the fixed charges, an attracting image-based force is imposed on each particle by the fixed charges, while a repelling particle-based force is imposed by the particles upon each other¹. These forces are computed respectively using (2). Their normalised weighted sum, reduced by a damping factor, plays the role of acceleration for each particle. As the image-based force has larger weight than the particle-based force, the particles primarily move towards the nearest and strongest edges. The repelling forces then try to advance the particles along the boundary until they have reached an equilibrium state, thus detecting the entire boundary. A multiscale approach was used to partially alleviate the heavy computational costs, and also to allow particles quickly spread across the image domain at coarser levels to capture as many boundaries as possible.

The CPM model [8] benefits from initialisation that is largely insensitive to placement. Nevertheless, it is computationally intensive as (a) particles have to advance along boundaries in order to encompass the desired object, and (b) particles are added and deleted dynamically at each iteration. Although a damping factor is used to reverse the direction of acceleration when a particle crosses an edge, the particle will still move as long as its speed is not exactly zero, and therefore oscillations occur at the boundaries and particle convergence needs to be flagged by some criterion. Above all, CPM can not guarantee closed contours, inevitably resulting in gaps in the recovered object boundaries particularly if the object is occluded or has weak edges. Furthermore, a final reconstruction of points into curves for continuous representation of object boundaries is necessary which may not encapsulate the true boundary of the object. Fig. 1 shows a synthetic image of a circle with a blurred edge region indicated by a black

¹ In [8], the attractive force is referred to as the Lorentz force in the absence of a magnetic field, and the repellent force as the Coulomb force.

window. In such regions the image-based forces are significantly influenced by the stronger edges nearby (see vector field in Fig. 1). As the image-based forces always dominate the direction of movement, particles which have arrived at the weak edges will continue moving to the stronger edges with the weak edges left unmarked. This leads CPM to fail to close the border around the synthetic circle.

3 Proposed Model: Charged Active Contour Model Based on Electrostatics (CACE)

The aim of our work is to improve on the drawbacks of the CPM particle model and the more traditional geometric contour models by integrating electrostatics principles with active contour evolution. Our proposed charged active contour model, CACE, detects objects starting with a *positively charged active contour* that propagates in an electrostatic field distributed with *negative fixed charges* proportional to image edge magnitudes.

The contour propagation in CACE results from the confluence of two components: a boundary attraction force and a boundary competition force. The attraction force acts as a bi-directional vector field which leads each point on the contour towards the boundaries from both sides. The competition force exerts most influence once any part of the snake reaches a boundary. It repels *free* contours nearby from reaching the already occupied boundary. The stronger the boundary, the larger the repelling force the contour exerts. This repelling force is also designed in a way such that only contours in homogeneous regions are most affected. In other words, contours that reach object boundaries will exert repellent forces upon other contours while they themselves will be least affected by others. At the same time, contours in homogeneous regions will continue to deform according to both attraction and competition forces. This is significantly different from the repelling force in the CPM model where the particles are constantly pushing each other in opposite directions. The electrostatic force field in the proposed CACE model is dynamically adapting as the contour evolves. This brings flexibility in initialisation and better curve propagation towards object boundaries. The CACE model is implemented in a geometric contour propagation framework using the Level Set representation to naturally handle topological changes.

We now describe in detail how these two forces are obtained and how they interact to create the joint electrostatic force field for the propagation of CACE.

3.1 Boundary Attraction Force Field

Let I denote an image and \mathbf{x} the pixel position. We use the Gaussian-based edge detector, with zero mean and variance σ_E , used in [7] as the boundary descriptor:

$$f(\mathbf{x}) = 1 - \frac{1}{\sqrt{2\pi}\sigma_E} \exp\left(-\frac{|\nabla(G_\sigma * I)(\mathbf{x})|^2}{2\sigma_E}\right), \quad \mathbf{x} \in \mathbf{X}. \quad (3)$$

where $G_\sigma * I$ denotes the convolution of the input image and a Gaussian smoothing kernel. The construction of the attraction force is based on the electrostatic

force interaction as given in (1) and (2). Here, we treat the boundary pixels, defined in (3), as fixed negative charges with magnitude proportional to their edge strength. Thus, given N as the number of negative charges at locations $\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N$ across the edge map then the negative charge assigned to each edge pixel \mathbf{r}_i , denoted as $q_{\mathbf{r}_i}$ and simplified to q_i , is $q_i = -f(\mathbf{r}_i) < 0$.

The electrostatic field $\mathbf{E}_A(\mathbf{x})$ generated by these negative fixed charges can then be computed according to (1) as:

$$\mathbf{E}_A(\mathbf{x}) = \sum_{i=1}^N \frac{q_i}{4\pi\epsilon_0} \frac{\mathbf{x} - \mathbf{r}_i}{|\mathbf{x} - \mathbf{r}_i|^3}, \quad \mathbf{x} \in \mathbf{X}. \quad (4)$$

This electrostatic vector field points towards the negative fixed charges, i.e. the edges, resulting in a bi-directional force field. The snakes can be hypothesised as positive charges moving in the image domain under the influence of the negative boundary charges with the aim of converging towards them from both sides. Let M be the number of positive charges at positions $\mathbf{s}_1, \mathbf{s}_2, \dots, \mathbf{s}_M$ on the contour, and $p_{\mathbf{s}_j}$, simplified to p_j , denote the positive charge assigned to point \mathbf{s}_j . The attractive electrostatic force \mathbf{F}_A enforced upon the contour is then:

$$\mathbf{F}_A(\mathbf{s}_j) = p_j \mathbf{E}_A(\mathbf{s}_j), \quad \mathbf{s}_j \in \mathbf{X}. \quad (5)$$

As the contour and the fixed charges have opposite polarity, the electrostatic boundary attraction force continuously pushes the contour towards object boundaries. In this study, a constant unit positive charge p_j is assigned to all snake points. However, p_j can be treated as a variable for other applications.

3.2 Boundary Competition Force Field

While the boundary attraction force is constantly pushing the snake towards boundaries, the boundary competition force allows progress only towards unoccupied object boundaries. It helps a snake already occupying an object boundary to check the advance of free contours nearby.

The competition force results from an electrostatic field which continuously adapts as the contour evolves and reaches boundaries. Two conditions characterise this force: (a) Contours that are on the object boundaries endow most to the electrostatic field with contributions proportional to the edge strength. (b) The force upon a contour due to this electrostatic field is proportional to the inverted strength of the edge covered by this contour. In other words, contours in homogeneous regions are most enforced upon while those on top of strong edges are least pushed. This ensures the snakes stay at their detected boundaries but push away nearby snakes competing for the same boundaries.

Condition (a) above is realised by weighting the contour charges with the edge function, i.e. $p'_j = f(\mathbf{s}_j)p_j$. The resulting electrostatic field comprises vectors pointing away from the edges already occupied by contours. It is given as:

$$\mathbf{E}_C(\mathbf{x}) = \sum_{j=1}^M \frac{p'_j}{4\pi\epsilon_0} \frac{\mathbf{x} - \mathbf{s}_j}{|\mathbf{x} - \mathbf{s}_j|^3} = \sum_{j=1}^M \frac{f(\mathbf{s}_j)p_j}{4\pi\epsilon_0} \frac{\mathbf{x} - \mathbf{s}_j}{|\mathbf{x} - \mathbf{s}_j|^3}, \quad \mathbf{x} \in \mathbf{X}. \quad (6)$$

Condition (b) is realised by weighting the contour charges with an edge stopping function, i.e. $g(\cdot) = 1 - f(\cdot)$, to generate the boundary competition force \mathbf{F}_C :

$$\mathbf{F}_C(\mathbf{s}_j) = g(\mathbf{s}_j)p_j\mathbf{E}_C(\mathbf{s}_j), \quad \mathbf{s}_j \in \mathbf{X}. \quad (7)$$

Thus, \mathbf{F}_C can be considered as a boundary competition force that prevents contours from approaching the same boundaries. For example, consider point charges p_a and p_b on the active contour at positions \mathbf{s}_a and \mathbf{s}_b . If these two points are both in homogenous regions, $\mathbf{E}_C(\mathbf{s}_a)$ and $\mathbf{E}_C(\mathbf{s}_b)$ are small, and they exert little competition force upon each other (and on other snake points). However, both of them are repelled by any other points that have already reached boundaries. When one of this pair, say p_a , reaches a boundary, it (along with all other snake points on object boundaries) will alter the electrostatic field according to (6), with its contribution to the field being proportional to its edge strength $f(\mathbf{s}_a)$. The impact of this electrostatic field on p_a itself is however minimised since the force $\mathbf{F}_C(\mathbf{s}_a)$ is weighted by $g(\mathbf{s}_a)$ (in (7)), i.e. the stopping function prevents it from being pushed away from the boundary. The snake point p_b on the other hand provides little contribution to this field, but will be most affected by the competition force $\mathbf{F}_C(\mathbf{s}_b)$ due to the large value of $g(\mathbf{s}_b)$ in the homogeneous region. When both snake points reach a boundary, they both contribute to the electrostatic field but have barely any influence on each other.

3.3 Joint Electrostatic Force

The joint electrostatic force \mathbf{J} on the active contour is obtained by combining (5) and (7) as such:

$$\begin{aligned} \mathbf{J}(\mathbf{s}_j) &= p_j[\lambda\mathbf{E}_A(\mathbf{s}_j) + (1 - \lambda)g(\mathbf{s}_j)\mathbf{E}_C(\mathbf{s}_j)] \\ &= \lambda\mathbf{F}_A(\mathbf{s}_j) + (1 - \lambda)\mathbf{F}_C(\mathbf{s}_j). \end{aligned} \quad (8)$$

The real positive constant λ balances the contribution between the boundary attraction force and the boundary competition force. As shown in sections 3.1 and 3.2, the first term attracts the contours to object boundaries, while the second term prevents the contours from approaching the boundaries that are already covered by other contours. The ever-changing force field causes the free contours to change direction and search for other boundaries.

It is important to further emphasise that the joint force field is dynamically adapting to the evolution of the snake and in turn defining its advance. The electrostatic attraction force field described in Section 3.1 is a *static* bi-directional vector field that attracts contours to object boundaries. A deformable contour model solely based on this static force field inevitably suffers from similar difficulties as the GVF snake model and its variations. Instead of attempting to overcome the saddle or divergent points in a vector field as proposed in [7], the CACE model adapts the vector field through the boundary competition force so that such critical points change as the snake approaches.

Fig. 2 illustrates adaptive changes of the joint electrostatic force field during contour propagation. The test image and the initial CACE snake are shown in

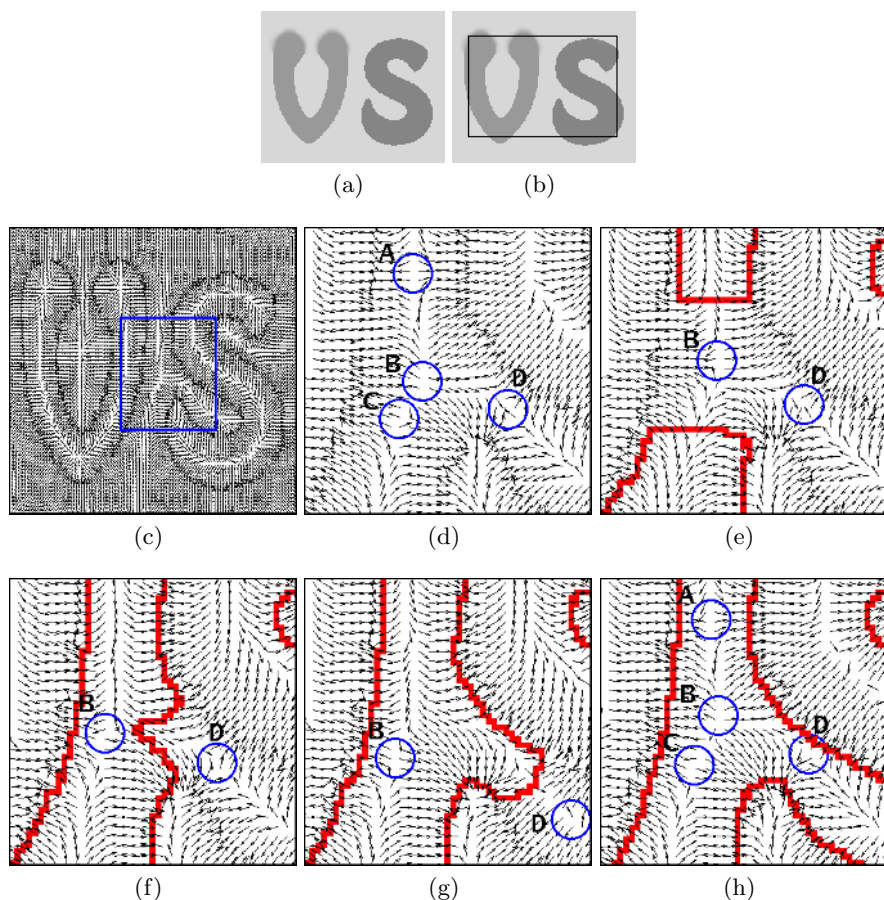


Fig. 2. Change of force fields during contour propagation. (a) Test image with letters 'V' and 'S'; (b) Initial snake; (c) Initial CACE vector field with marked region; (d) Closeup view of the vectors in the marked region in the valley between 'V' and 'S'; (e)-(f) Adapting vector field as snake progresses. Snake positions are indicated in thick dark red, and critical points are shown in thin blue circles.

2(a) and 2(b) respectively. Fig. 2(c) shows the initial vector force field and Fig. 2(d) is a closeup of the square region marked in 2(c) in the valley between the letters 'V' and 'S'. There are four critical points, indicated by thin blue circles, that can stop the snake from further propagation. A, C, and D are saddle points, while B is a divergent point. The thick red contours in Figs. 2(e)-2(h) are the progressing positions of the CACE snake. In 2(e), as the snake evolves in the valley, the saddle points A and C disappear. Notably, the divergent point B becomes a saddle point. Saddle point D stays approximately the same, as the snake is still far away from it. In 2(f), the snake has just passed the valley and is going to enter the deep concave in the letter 'S'. In 2(g) the saddle point D is

clearly moving away from the entrance of the concavity as the snake approaches. Finally in 2(h), the snake reaches the boundaries and the vector field takes a similar form as the initial state. The saddle points A and C re-emerge, saddle point D is back to the entrance of the concave, and B changes back into a divergent point. The corresponding CACE evolutions are shown in the last row of Fig. 4.

3.4 Geometric Active Contour Formulation for CACE

Let C be the active contour. The contour evolution formulation for the CACE model is defined as:

$$C_t = \alpha g \kappa \mathcal{N} + (1 - \alpha)(\mathbf{J} \cdot \mathcal{N})\mathcal{N}, \quad (9)$$

where α is a real constant, κ denotes the curvature, and \mathcal{N} is the unit inward normal. The first term regulates the contour, and the second term attracts the snake towards the object boundaries. To ensure efficient contour propagation, we normalize the force field along the contour normal by replacing the term $(\mathbf{J} \cdot \mathcal{N})\mathcal{N}$ with $\frac{(\mathbf{J} \cdot \mathcal{N})\mathcal{N}}{|(\mathbf{J} \cdot \mathcal{N})\mathcal{N}|}$.

To achieve topological flexibility, we use level sets [9] to represent the contour, implicitly evolving it by deforming the level set function, u . This involves two extensions. The first is to embed the 2D contour into a 3D level set function u , which is achieved by using the signed distance transform such that the embedded snake is given by the zero level set at any time. The second is to extend the force field defined on the 2D contour to the 3D level sets. The Fast Marching Method can be used to accomplish this as proposed in [10]. However, in this study, we can simply compute the extended force field by treating each level set as a deforming contour at each time step. Thus, the joint force field $\mathbf{J}(\mathbf{s}_j)$ as given in (9) is extended to $\mathbf{J}(\mathbf{x})$ across the image domain. Thus, given the fact that $\mathcal{N} = -\frac{\nabla u}{|\nabla u|}$, the level set representation of our CACE snake is given as:

$$u_t = \alpha g \kappa |\nabla u| - (1 - \alpha)\mathbf{J} \cdot \nabla u. \quad (10)$$

4 Experimental Results

In this section we present results for our CACE model and compare its performance against the CPM, the geodesic contour, and the GVF geodesic contour models. The software for all the methods we compare against was developed in-house based on the relevant literature, i.e. [8, 7, 11].

CACE copes much better than CPM when faced with weak edges (cf. Fig. 3 with Fig. 1). As CPM particles arrive at weak edges, they carry on moving towards stronger edges along the boundaries, hence fail to correctly recover the object boundaries. CACE stabilizes around the boundaries, successfully detecting the whole object, due to the bi-directional nature of its force field and the characteristics of the contour itself. The vectors pointing towards the edges, although weak, prevent leakage from both sides.

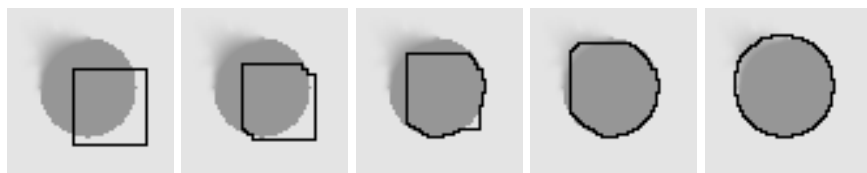


Fig. 3. Propagation of CACE on disc object with weak edges (cf. with Fig. 1)

CACE possesses significant advantages over other contour models, e.g. it is more robust to initial placement than the geodesic snake, and better capable of handling object topology than the GVF geodesic snake, as shown in Fig. 4. While the geodesic snake fails to detect the objects under initialization that crosses boundaries, the GVF geodesic snake is less constrained, but nonetheless, still unable to reach some of the boundaries when it gets trapped by divergent vectors in homogeneous areas. CACE improves on these limitations and succeeds in detecting both objects in Fig. 4.

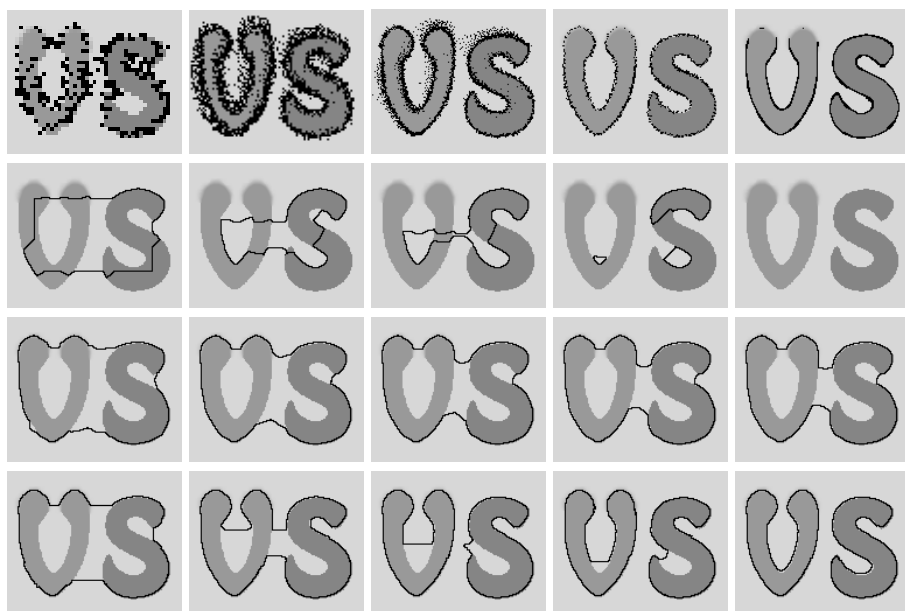


Fig. 4. Contour propagation for boundary detection. Top row: iterations of the CPM, 2nd row: geodesic snake, 3rd row: GVF geodesic snake, final row: CACE.

Fig. 5 shows the evolution process in CPM, geodesic snake, GVF geodesic snake, and CACE, on a corpus callosum detection task in an MRI brain image. Although the coarse-to-fine multi-scale setting is used, CPM still fails to

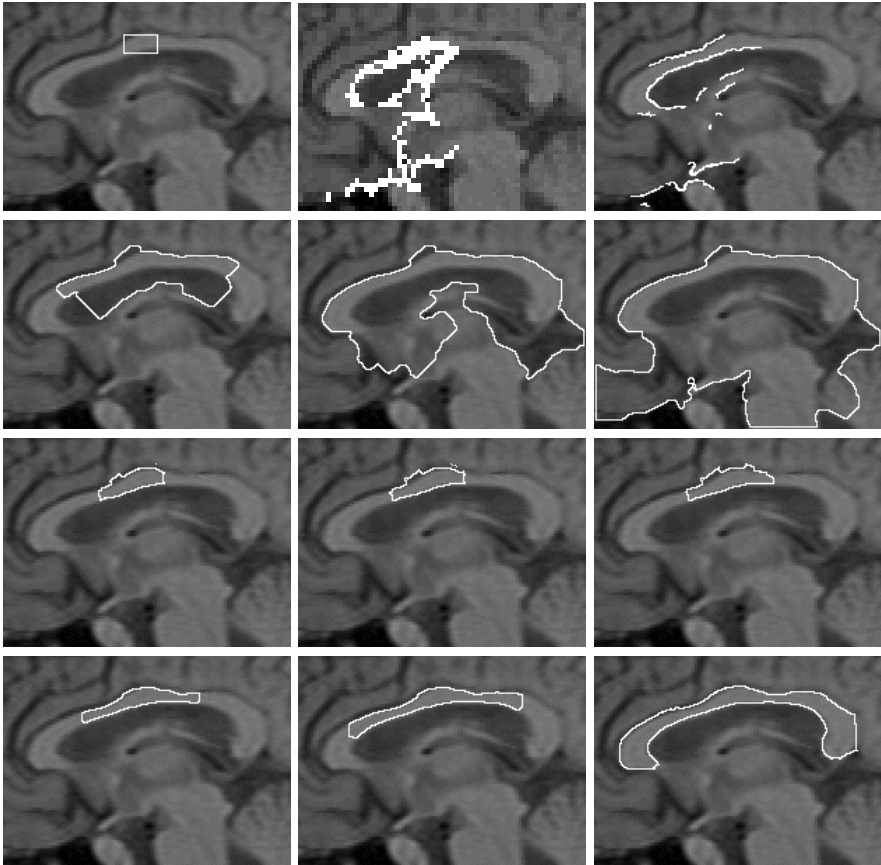


Fig. 5. From left to right, top: input image and initialisation for all the models, iteration of the CPM, final CPM result, 2nd row: evolution of the geodesic snake, 3rd row: evolution of the GVF geodesic snake, final row: evolution of CACE

recover the corpus callosum, as particles can not advance from stronger boundaries towards weaker ones and are thus blocked in the area where strong edges occur. The geodesic snake also fails in the detection task due to initialization across boundaries, as does the GVF geodesic snake which gets trapped by saddle points formed within the corpus callosum. In comparison, CACE benefits from the self-adaptive nature of the force field and manages to propagate through the elongated part of the object and capture the entire boundary.

Figs. 6 and 7 show more examples where CACE again performs more accurately than the other snake models under highly noisy and textured conditions.

The CACE model performs well on a range of parameter settings. Two main parameters are involved: (λ, α) . The parameter λ in (9) balances the contribution between the attraction and the competition forces. We set $\lambda = 0.3$ throughout our experiments determined empirically for the set of images shown. The pa-

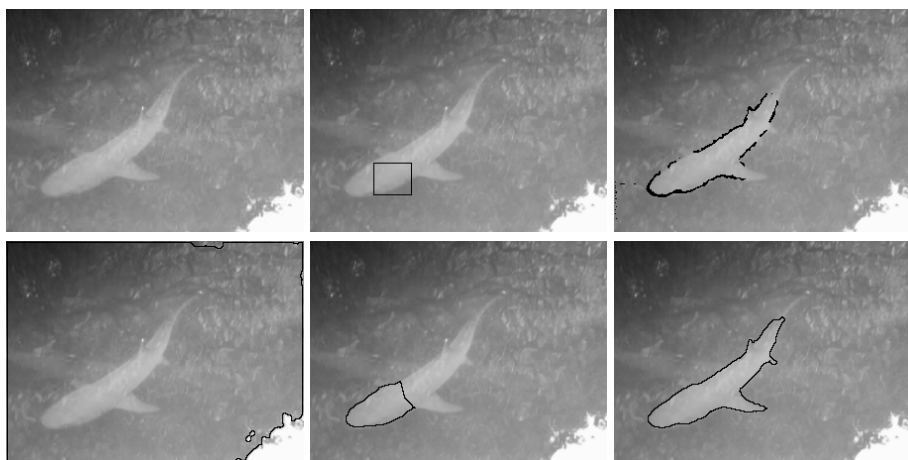


Fig. 6. From left to right, top: noisy input image, initialization for all models, CPM results, and bottom: results for the geodesic snake, the GVF geodesic snake, and CACE

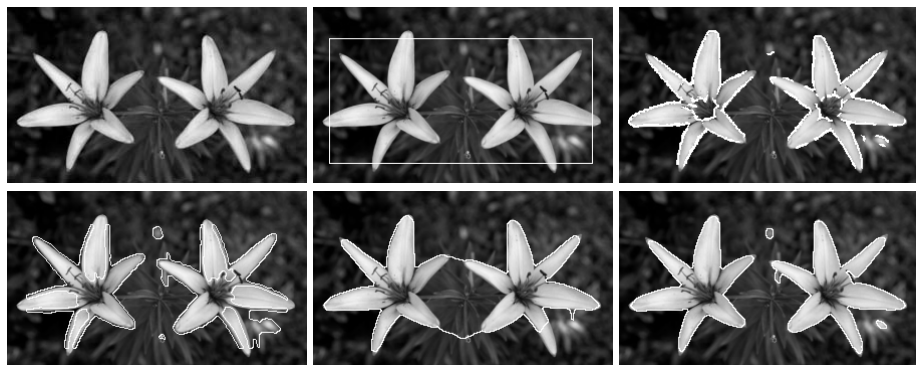


Fig. 7. From left to right, top: input image, initialization for all models, CPM results, and bottom: results for the geodesic snake, the GVF geodesic snake, and CACE

parameter α controls the smoothness of the contour and has minor impact on the model performance and was kept constant at $\alpha = 0.1$ throughout our work.

It is worth noting that the computation of the electrostatic force field in (4) or (6) is simple but inefficient, requiring $O(N^2)$ computational complexity and increases drastically as the image size increases. Therefore, as with CPM in [8], we use the Particle-Particle Particle-Mesh method, originally proposed in [12], for fast and accurate evaluation of the electrostatic field. Details of the method can be found in [12]. In terms of comparative computational performance, we used a 200×200 image in which all models successfully found the object. Using

a 2.8 GHz Linux PC running uncompiled Matlab code, the computational times for the different particle and contour models were as follows: 281s for CPM, 26s for the geodesic snake, 20s for the GVF geodesic snake, and 29s for CACE.

5 Conclusion

In this paper, we presented a novel active contour model, namely the Charged Active Contour, CACE. It incorporates electrostatics principles from the CPM particle model [8] into the deformable contour model. The CACE snake deforms under the confluence of an external boundary attraction force and an external boundary competition force. Driven by this combined electrostatic force, contours move towards object boundaries, and will end up there if the boundaries are not covered by other contours, or change direction and search for other boundaries otherwise.

Experimental results have demonstrated that by introducing particle dynamics into the contour model, the snake can be more initialisation independent, exhibit better ability in reaching concavities, and ensure closed contours.

References

1. Kass, P., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. *IJCV* **1** (1988) 321–331
2. Caselles, V., Kimmel, R., Sapiro, G.: Geodesic active contours. In: *ICCV*. (1995) 694–699
3. Zhu, S., Yuille, A.: Region competition: Unifying snakes, region growing, and Bayes/MDL for multiband image segmentation. *IEEE T-PAMI* **18** (1996) 884–900
4. Paragios, N., Deriche, R.: Geodesic active regions for supervised texture segmentation. In: *ICCV*. (1999) 926–932
5. Xie, X., Mirmehdi, M.: RAGS: Region-aided geometric snake. *IEEE T-IP* **13** (2004) 640–652
6. Xu, C., Prince, J.: Gradient vector flow: a new external force for snakes. In: *CVPR*. (1997) 66–71
7. Paragios, N., Mellina-Gottardo, O., Ramesh, V.: Gradient vector flow fast geodesic active contours. *IEEE T-PAMI* **26** (2004) 402–407
8. Jalba, A., Wilkinson, M., Roerdink, J.: CPM: A deformable model for shape recovery and segmentation based on charged particles. *IEEE T-PAMI* **26** (2004) 1320–1335
9. Sethian, J.: *Level Set Methods: Evolving Interfaces in Geometry, Fluid Mechanics, Computer Vision, and Materials Science*. CUP (1996)
10. Adalsteinsson, D., Sethian, J.: The fast construction of extension velocities in level set methods. *J. Comp. Phy.* **148** (1998) 2–22
11. Caselles, V., Catta, F., Coll, T., Dibos, F.: A geometric model for active contours. *Numerische Mathematik* **66** (1993) 1–31
12. Hockney, R., Eastwood, J.: *Computer Simulation Using Particles*. Taylor and Francis (1988)

Comparison of Statistical and Shape-Based Approaches for Non-rigid Motion Tracking with Missing Data Using a Particle Filter

Abir El Abed, Séverine Dubuisson, and Dominique Béréziat

Université Pierre et Marie Curie - Laboratoire d'Informatique de Paris 6
8 rue de Capitaine Scott, 75015 Paris, France
{abir.elabed, severine.dubuisson, dominique.bereziate}@lip6.fr

Abstract. Recent developments in dynamic contour tracking in video sequences are based on prediction using dynamical models. The parameters of these models are fixed by learning the dynamics from a training set to represent plausible motions, such as constant velocity or critically damped oscillations. Thus, a problem arise in cases of non-constant velocity and unknown interframe motion, *i.e.* unlearned motions, and the CONDENSATION algorithm fails to track the dynamic contour. The main contribution of this work is to propose an adaptative dynamical model which parameters are based on non-linear/non-gaussian observation models. We study two different approaches, one statistical and one shape-based, to estimate the deformation of an object and track complex dynamics without learning from a training set neather the dynamical nor the deformation models and under the constraints of missing data, non-linear deformation and unknown interframe motion. The developed approaches have been successfully tested on several sequences.

1 Introduction

Many problems require the estimation of the state process of a dynamic system using a sequence of noisy measurements. Filtering is used in a widely applications, such as tracking problems in image processing. An optimal solution to predict the state process is given by the bayesian approach which aim is to construct the posterior probability density function of the state using all available informations. When the system and measurements are linear with gaussian additive noise, the density is characterized by its mean and covariance matrix. In image processing, the optimal solution requires a relatively long computation time and cannot be solved, except for gaussian linear systems: Kalman filter provides an analytical recursive expression for the first two moments [1]. For non-linear systems, extensions of Kalman filter have been developed [2], but the performance is quickly degraded with time when the system presents a strong nonlinearity. Thus, when the model is highly nonlinear and/or non gaussian, Kalman filter and its extensions fail to give an accurate approximation of the mean and the covariance matrix, and make the problem of optimal filtering difficult to solve. Under such difficulties, numerical methods, such as the Sequential

Monte Carlo methods, are particularly appropriate to approximate the posterior probability density function of the state. These approaches are known as particle filters and mainly consist in propagating a weighted set of particles that approximates the density function. They provide flexible tracking frameworks as they are limited neither to linear systems nor to Gaussian noise [3,4,5].

According to Isard and Blake in [6,7], dynamic contour tracking is based on predictions using dynamical models. The parameters of these models are fixed by hand to represent plausible motions, such as constant velocity or critically damped oscillations. Experimentations allow these parameters to be refined by hand to improve tracking but this is a difficult and unsystematic business, especially in a high-dimensional shape-space which may have complex couplings between the dimensions. It is far more attractive to learn dynamical models on the basis of training sets. Once a new dynamical model has been learned, it can be used to build more efficient trackers. It can more accurately track the original training sequence or a new testing sequence, involving greater agility of motion. In practice, they incorporate the learned model into the CONDENSATION algorithm [8], estimation process which should enable particles to be concentrated more efficiently. This allows the curve motion to be estimated correctly with N particles in each time step. In this framework, learning the dynamics is required to achieve and succeed the task of tracking. Although, it may fail if the motion is not anticipated by the learned model. To resume, the performance of a building tracker is based on the parametrisation of the dynamic model.

In practical interpretation problems, the complex dynamics, such as non-constant velocity or non-periodic oscillations, make too difficult the choice of the parameters of the dynamical model for an estimation algorithm. Furthermore, the learning step becomes particularly more difficult in case of missing data where we don't know neather the dynamical nor the deformation models in an interval of time between two successive observations (see Figure 1). Adaptative and automated parameters of dynamics is of crucial importance. Therefore, as the motion has non-constant velocity and/or non-periodic ocillations, dynamical model parameters are needed in order to determine the settings of the estimation parameters. In addition, in the case of deformable curves, the probabilistic framework is not sufficient to track them and estimate the parameters of their affine transformations (*i.e.* translation, rotation and scale).

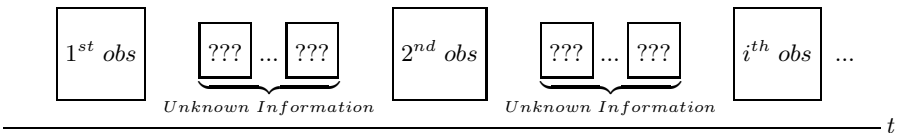


Fig. 1. A set of observations: we don't know neather the dynamical nor the deformation models in the interval of time between two successive observations

In this paper, we propose to model the dynamics in an adaptative way, without any learning from a training sequence, only by using the available measurements.

In addition, we suggest two approaches, a statistical one and a shape-based one, for tracking non-rigid contour in a video sequence under the constraints of unknown interframe motion, missing data, non-linear/non-gaussian observations and non linear dynamics. Our aim rest on the estimation of the affine transformation and the dynamics of the discrete contour of an object between two observations occurred at times t and $t + \delta t$. The representation of a discrete contour by pixels is impossible since their number is too high. Moreover, some pixels are redundant for the smooth portions whereas other more irregular portions are represented inaccurately. In this case, we need an abstraction of the contour which ideally is invariant against translations, scales, rotations, and starting point, while still representing the essential form of the contour. We then propose a method based on a frequency-domain decomposition where the discrete contour is represented by the Fourier descriptors. This representation consists in breaking up the contour according to a base of orthonormal functions and can moreover truncate the parameters of high frequencies. The main advantage of these descriptors is that they contain the parameters of the affine transformation, (*i.e.* translation, rotation and scale), undergone by the contour from one frame to another one. We propose a shape-based approach which is divided into several steps. We first compute the Fourier descriptors of the contour via a Discrete Fourier Transform (DFT), then make a high-pass filtering and perform an Inverse Discrete Fourier Transform (IDFT) to reconstruct the truncated contour. These remaining points are tracked using a particle filter so that we can extract, from the estimated Fourier descriptors, the motion and the affine transformation of the contour.

The idea of the statistical approach consists in estimating the underlying motion only by using the statistical parameters of an object: mean and variance. Using the available observations and particle filter, we estimate its mean in the interval of time that separate two successive observations. Then, we interpolate its variance using a cubic B-Spline to measure the spread around its mean and compute its orientation using its variance.

The outline of this paper is as follows. Section 2 presents a brief review on particle filter. The shape-based and statistical approaches are respectively developed in Sections 3 and 4. In Section 5, we explain the adaptative dynamical model and test the previous approaches on different video sequences to show their robustness. Finally, comparisons and conclusions are given in Sections 6 and 7.

2 Particle Filter

Given a video sequence depicting a moving object, the tracking consists in estimating its state process X_t in frame t . Particle filtering, or Sequential Monte Carlo algorithm, is an inference process which can be considered as a generalization of the Kalman filter [3,4]. It aims at estimating the unknown state X_t from a set of noisy observations that occurred sequentially, $Y_{1:t} = (y_1, \dots, y_t)$. Two important components of this approach are the state transition and observation models whose most general forms can be given by $X_t = F_t(X_{t-1}, U_t)$

and $Y_t = G_t(X_t, V_t)$, respectively. We notice that U_t is the system noise, F_t the kinematics, V_t the observation noise, and G_t the observation model. The particle filter approximates the posterior distribution $P(X_t|Y_{1:t})$ by a set of weighted particles $\{s_t^{(j)}, w_t^{(j)}\}_{j=1,\dots,N}$. This set of particles represents the state of the object and $w_t^{(j)}$ is the discrete probability of the particle $s_t^{(j)}$. Generally, the displacement of the particles is computed from an appropriate density f which depends on the available data [9]. The particle filter algorithm proceeds as follows:

Initializing generate the particle set $S_0 = (s_0^n, w_0^n)$ where $s_0^n \sim P(X_0)$ and $w_0^{(i)} = \frac{1}{N}$;

Resampling \tilde{s}_t^j from $P(X_t|X_{t-1} = s_{t-1}^j, Y_t = y_t)$;

Weighting and normalizing $\tilde{w}_t^j = w_{t-1}^j \frac{P(\tilde{s}_t^j|s_{t-1}^j)P(y_t|\tilde{s}_t^j)}{f(\tilde{s}_t^j|s_{t-1}^j, y_t)}$ and $w_t^j = \frac{\tilde{w}_t^j}{\sum_{j=1}^N \tilde{w}_t^j}$.

where $n, j = 1, \dots, N$ and $t = 1, \dots, T$. Finally, the density function can be approximated by $\sum_{n=1}^N w_t^n \delta_{s_t^n}$. It has been shown that, after few iterations, the variance of the particle weights always increases over time, which causes the weight degeneracy phenomenon [9]. A resampling of the particle weights is then required to reduce this effect. Resampling consists in selecting, during filtering, samples with high weights while those with relatively low weights are not. Using the multinomial resampling approach [10], we resample the particles in an adaptive way when their effective number is estimated by seeking a value for $N_{\text{eff}} = \frac{1}{\sum_{n=1}^N (w_t^n)^2}$ that is under a given threshold.

3 Shape-Based Approach

To track a deformable object under the constraints of unknown interframe motion and non linear dynamics, we suggest an approach based on a combination of a frequency-domain decomposition of the contour and a particle filtering using an adaptative dynamical model. Our idea for tracking a dynamic contour and estimating its deformation amounts to:

- Compute, for each observation, the Fourier descriptors of the contour via DFT;
- Smooth the contour by removing the harmonics of high frequencies (e.g low-pass filtering) and only keeping the r first Fourier descriptors;
- Compute the IDFT of the remaining descriptors to reconstruct the truncated contour with the remaining $(2r + 1)$ descriptors;
- Track the remaining points between frames t and $t + \delta t$ using a particle filter and compute their Fourier descriptors via DFT. Then, extract from these descriptors the parameters of the affine transformation, by comparison with the Fourier descriptors computed in frame t , and rebuild the contour only using $(2r + 1)$ descriptors.

3.1 Fourier Descriptors

Consider the N points, $P_i(x_i, y_i)_{i=1\dots N}$, of a contour as a discrete function $C = (x, y)$ where $x = (x_1, x_2, \dots, x_N)^t$ and $y = (y_1, y_2, \dots, y_N)^t$. We can describe C

in the frequential domain as a discrete complex function $U = X + jY$ with $U(i)$ its i^{th} component. The result can be transformed back into the spatial domain via IDFT without any loss. DFT and IDFT are defined respectively by $U(k)$ and $c(n)$: $U(k) = \sum_{n=0}^{N-1} c(n)e^{-\frac{2\pi j}{N}kn}$ and $c(n) = \frac{1}{N} \sum_{k=0}^{N-1} U(k)e^{\frac{2\pi j}{N}kn}$, where $(-\frac{N}{2} \leq k, n \leq \frac{N}{2} - 1)$. The coefficients $U(k)$ are also called Fourier descriptors [11]. They describe the discrete contour of an object in the Fourier domain. Some geometrical transformations of the contour function $U(k)$ can be related to simple operations in the Fourier domain. Translation by k_0 only affects the first Fourier descriptor $U(0)$. Scale of the edge with a factor α leads to scaling all the Fourier descriptors by α . Rotating the edge by an angle θ_0 yields a constant phase shift of θ_0 of the Fourier descriptors. Changing the starting point of the edge to the position n_0 results in a linear phase shift of $\frac{2\pi n_0 k}{N}$ of the Fourier descriptors.

The truncated contour obtained by applying a IDFT after removing the harmonics of high frequencies (e.g keeping the r first Fourier descriptors) is given by: $C_v^r = a(0) + \sum_{k=1}^r a(k) \exp(\mu) + \sum_{k=1}^r a(N-k) \exp(-\mu)$, where $v = (0, \dots, N-1)$, $\mu = 2j\pi \frac{kv}{N}$. The quantity of lost information if we only keep r coefficients can be measured by the following quadratic error: $E_N^r = \frac{\sum_{k=r+1}^{N-r-1} |a(k)|^2}{\sum_{k=1}^{N-1} |a(k)|^2}$.

The number r of coefficients to be preserved is determined such as the error E_N^r is lower than a given threshold β . We have adopted this modeling to rebuild the contour of the tennis man in Figure 2.a only using $(2r + 1)$ descriptors. The threshold β is fixed to 0.05 and Figure 2.b shows that $E_N^r \leq 0.05$ for $r = 2$ for the first frame and $r = 3$ for the second frame. Using the Fourier descriptors, we can extract the parameters of the affine transformation undergone by the tennis man from frame 1 to frame 45. The translation vector is $(-2, 61)$ pixels in the (x, y) plane, the scaling factor is 1.35 and the rotation is -8° . Figure 2.c shows the rebuilt contour for the tennis man using different values of r .

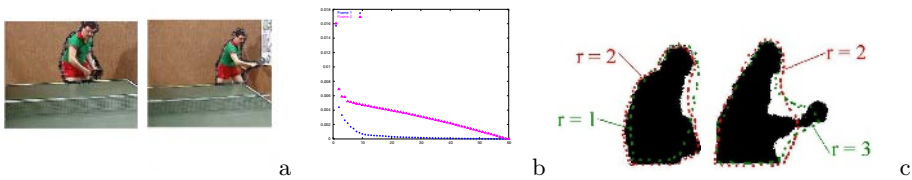


Fig. 2. Test in Tennis sequence:(a) Frames 1 and 45; (b) Error evaluation for both frames according to r ;(c) Rebuilt contour of the tennis-man using different values of r

4 Statistical Approach

To track an object and estimate its deformation between two observations, we suggest to estimate the parameters of its affine transformation only by using its statistical parameters (mean and variance). The main steps of this approach are given below:

- We apply the particle filter to estimate the mean $\hat{X}_t^i = \sum_{n=1}^N w_t^n \mathfrak{z}_t^{n,i}$ which represents the translation vector of the object;
- We interpolate the variance (to measure the spread around the mean) using a cubic B-spline, whose control points are the variance extracted from the available observations. The cubic B-splines are parameterized curves defined by a sum of N_B basic functions: $c(u) = \sum_{i=0}^{N_B-1} P_i B_i^3(u)$, where P_i is the i^{th} control point and N_B the number of control points. The B_i^k are polynomial functions of degree k and recursively defined using the selected nodes $(t_i)_{i=0, \dots, N_B+k-1}$:

$$B_i^1(u) = \begin{cases} 1 & \text{if } t_i \leq u \leq t_{i+1} \\ 0 & \text{else} \end{cases} \tag{1}$$

$$B_i^k(u) = \frac{(u - t_i)B_i^{k-1}(u)}{t_{i+k-1} - t_i} + \frac{(t_{i+k} - u)B_{i+1}^{k-1}(u)}{t_{i+k} - t_{i+1}} \tag{2}$$



Fig. 3. Test on "Taxi" sequence: estimation of the orientation of the white taxi (red arrows in both frames)

- We define the orientation of an object in the plane (x, y) by the slope m of the following equation:

$$y = mx + b = \frac{\sigma_y}{\sigma_x}x + \frac{\sigma_x \times \mu_y - \sigma_y \times \mu_x}{\sigma_x} \tag{3}$$

where (μ_x, μ_y) and (σ_x^2, σ_y^2) are respectively its mean and variance. As an example, we suggest to define the orientation of the white taxi shown in Figure 3. We compute the mean μ and the standard deviation σ in both frames: the numerical results are approximately $(\mu_{x_1}, \mu_{y_1}) = (562, 463)$, $(\sigma_{x_1}, \sigma_{y_1}) = (51, 25)$, $(\mu_{x_2}, \mu_{y_2}) = (504, 422)$ and $(\sigma_{x_2}, \sigma_{y_2}) = (34, 25)$. We then compute the orientation (see Eq. 3) that is symbolized by arrows in Figure 3. As we can see the orientation of the taxi is well detected in both frames.

5 Experimental Results

The modeling of the tracking problem consists in defining the state vector of the object $X_t = (x_t, y_t, v_{x_t}, v_{y_t}, a_{x_t}, a_{y_t})$, implying position, velocity and acceleration. The observation model is only defined by a position $Y_t^i = (x_t, y_t)$. We have chosen to characterise the dynamical model with a third order equation:

$$x_t = c_1(t)x_{t-1}^3 + c_2(t)x_{t-1}^2 + c_3(t)x_{t-1} + c_4(t) \tag{4}$$

where $(c_1(t), c_2(t), c_3(t), c_4(t))$ are adaptative parameters that vary with time according to the nature of the motion: their values are selected from the available observations. In addition, the velocity and acceleration are respectively given by: $v_{x_t} = \dot{x}_t$ and $a_{x_t} = \ddot{x}_t$. The associated discretized state equation, with time period Δ_t , is given by:

$$X_{t+\Delta_t}^i = \begin{pmatrix} I_d & I_d & I_d \\ 0 & I_d & 2 \times I_d \\ 0 & 0 & I_d \end{pmatrix} X_t^i + \begin{pmatrix} \frac{\Delta_t^2}{2} I_d \\ \Delta_t I_d \\ I_d \end{pmatrix} U_t + W_t \tag{5}$$

where I_d is the 2×2 identity matrix, W_t a non linear function depending on the variables $(c_1(t), c_2(t), c_3(t), c_4(t))$ and U_t is a Gaussian zero-mean vector with covariance matrix $\Sigma_U = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix}$.

5.1 Cases Study

For all the following tests, we use the same dynamic noise ($\sigma_x = \sigma_y = 0.01$) to predict particles and the number of particles is set to 100. The threshold of the resampling step is equal to 0.3. Also, for all the tests, we plot the Regression Error Characteristic (REC) curve [12], the error rate is on the x-axis and the accuracy is on the y-axis. Accuracy is defined as the percentage of points that are fit within the tolerance. If we have a zero tolerance, we only consider the points that the function fits exactly as accurate. If we choose a tolerance that exceeds the maximum error observed for the model on all the data, then all points would be considered accurate. The error here is defined as the difference between the actual value and its prediction for any point (x, y) . As the error increases, the accuracy increases. The accuracy goes to 1 when the error becomes large enough.

a. Tennis table’s ball tracking: the more difficult problem when tracking the ball of a tennis table is that the motion is oscillatory and of a duration that is not an integer multiple of the period of oscillation (see Figure 4.(a-b)). As shown in Figure 4.b, the dynamics of the ball is complex and undergoes vertical and horizontal oscillations with different periods coupled with translation in both direction. Furthermore, the velocity is non-constant and the movement of the ball accelerates and decelerates according to the blow given by the player. In such kind of systems, it is very difficult, even impossible, to learn the motion from a training set because it is non-linear, non-periodic and variable from a sequence to another. For this reason, we propose the use of an adaptative dynamical model (Eq. 5) which parameters vary in time and depend on the available observations. Figure 4.a represents some observation frames showing the vertical and horizontal oscillations of the ball. We notice that the observations are not ocured with regular times. To track the ball of Figure 4.a without learning its dynamic, by only using 15 observations (in a sequence of 90 frames), we incorporate the adaptative dynamical model into the particle filter. The dotted line in Figure 4.b shows the estimated trajectory of the ball. Despite the high non linearity of the dynamic and the constraint of missing data, we have

obtained 8.13% as an average estimation error for the position which prove the effectiveness of the adaptative dynamical model. Figure 4.c shows the REC curve (error tolerance on the x-axis versus the percentage of points predicted within the tolerance on the y-axis).

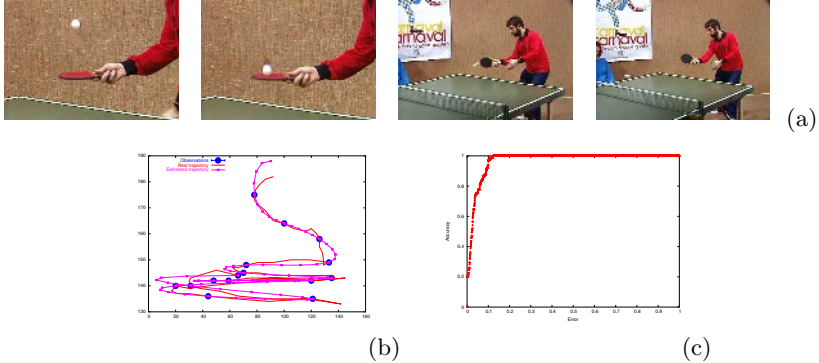


Fig. 4. Test on Tennis Table sequence: (a) Some observation frames. From left to right {1, 11, 50, 71 and 78}; (b) The dotted and solid lines respectively represent the estimated and real trajectory. The blue points represent the observations; (c) REC curve plots the error rate on the x-axis and the accuracy on the y-axis.

b. Bio-cellular tracking: we have tested our approaches on biocellular video sequences where the motion of a dark structure evolving on the surface of a cell is non rigid (see Figure 5.c). We just use an observation every 10 frames and we do not have any prior knowledge about the motion and the deformation of the dark structure between two successive observations. We suppose that this structure is well segmented. Using the shape-based approach, the state vector is composed of $(2r + 1)$ Fourier descriptors. For the sequence of Figure 5.c, the error is smaller than 5% for $r = 2$. Five descriptors are sufficient to track the motion of the dark structure with the particle filter and estimate its deformation. Figure 5.a shows the position estimated with the particle filter for each point and Figure 5.c shows that the estimated structure is closed to the original one. Figure 5.b shows the REC curve which plots the error tolerance on the x-axis versus the percentage of points predicted within the tolerance on the y-axis. Using the statistical approach, we track the mean of the structure using the particle filter and then can estimate its translation vector between two observations (Figure 5.e). The variance is then computed by

Table 1. Error ratio between the original and the estimated structure with both approaches

Frames	3	7	12	15	18
Shape-Based	3.2	3	4.5	5.35	4.9
Statistical	0.8	1.05	2.75	2.18	1.7

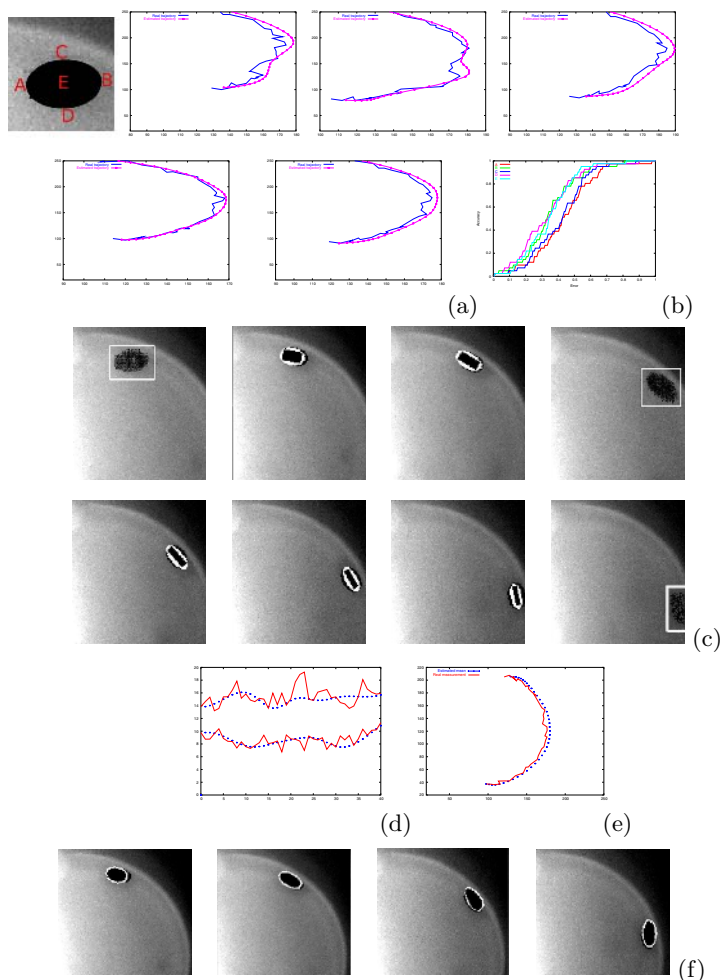


Fig. 5. Test on Bio-cellular sequences: (a) Estimated trajectory of points $\{A,B,C,D,E\}$ by using the particle filter; (b) REC curve: error rate on the x-axis and the accuracy on the y-axis for the predicted positions; (c) Results of tracking using the shape-based approach where the estimated structure (white contour) is close to the original (black structure). The frames of this sequence are numbered as follows: $\{1, 3, 7, 10, 12, 15, 18, 20\}$, and the frames $\{1, 10, 20\}$ are the observations. In this sequence the observations are squared; (d) Dotted and solid lines are the interpolated and real variance; (e) Dotted and solid lines are the mean estimated with particle filter and real measurements; (f) Results of tracking using the statistical approach: the estimated structure (white contour) is close to the original (black structure).

interpolation using a cubic B-Spline whose control points are the variance extracted from the observations (Figure 5.d). Finally, the estimation of the orientation is done by using the estimated variance (Section 4). The ratio error

of the estimated mean and variance is 3.8% and (1.73%, 2.6%), respectively. Figure 5.f shows that the estimated structure is closed to the original, that confirms the effectiveness of our approach. The ratio of loss information in both approaches is given in Table 1.

c. Ants Tracking: we have tested the statistical approach on a video sequence showing ants whose motion is unknown and highly non linear: the solid lines in Figure 6.a show their trajectories.

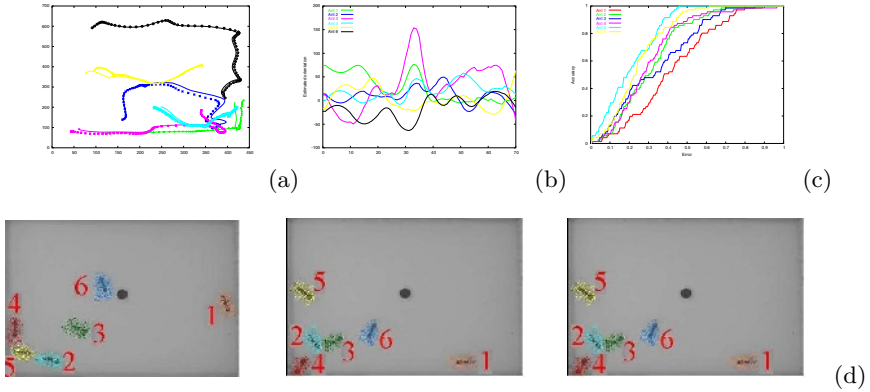


Fig. 6. Test on " Ants " sequence: (a) Dotted lines: the estimated mean of the ants; solid lines: the real mean; (b) The estimated orientation of the Ants; (c) REC curve: error rate on the x-axis and the accuracy on the y-axis; (d) The tracking result obtained with the statistical approach

To track these ants under the assumption of missing data, we only use five frames from the sequence as observations without having any prior knowledge about the motion or the trajectory in the interval of time that separates two successive observations. The motion of these ants is sometimes limited to a rotation around their axis and sometimes is coupled with a translation vector. In another word, they move with a non-constant velocity and can accelerate, decelerate and sometimes stop moving or starting. Our goal is to estimate the components of their motion (translation and orientation), only using their mean and variance. We apply the particle filter to estimate their mean, $\hat{X}_t^i = \sum_{n=1}^N w_t^n \tilde{z}_t^{n,i}$, which represents their translation vector. The dotted lines in Figure 6.a show the estimated mean for each ant. The orientation of an object is defined by its mean and variance (Section 4). Thus, under the assumption of missing data, the variance of the ants is only known in the available observations. We notice that some part of the ant's body can rotate which implies a variation of the variance. The variance of each ant is computed by interpolation using a cubic B-Spline method whose control points are the given variance extracted from the observations. After the estimation of mean and variance of the ant, we generate a set of random points, having the same estimated mean and variance, to localize its position and orientation. Figure 6.c shows for each ant the REC (the error rate on the x-axis versus the percentage of points predicted within the tolerance on the y-axis). Figure 6.b shows

the estimated orientation for each ant which are computed from their interpolated standard deviation (see Eq. 3). Figure 6.d shows that the generated set is closed to the ant, which confirms the effectiveness of our approach.

d. Tennis Man Tracking: to track the dynamics and estimate the deformation of the tennis man, we use the shape-based approach. We smooth the contour of the player after removing the harmonics of high frequencies. The number r of coefficients to be preserved is 6 for an error $\leq 5\%$. In figure 7, at the beginning the curve gives a rough approximation of the shape of the player, but progressively in the sequence this one really sticks to the shape of the player. Our aim is to estimate the evaluation of his deformation/appearance with time only using four observations. The black shape in Figure 7 represents the estimated shapes between observations while the color ones are the observations. We represent the remaining points to be tracked using the particle filter with the green arrows (last image in Figure 7).



Fig. 7. Test on "Tennis Man" sequence: the green arrows (see last image) point on the remaining points to be tracked using the particle filter ($r = 6$ for an error $\leq 5\%$). The black structure are the estimated contour between observations, and the color image are the observations.

6 Discussions and Comparison of the Approaches

The purpose of this paper is to compare two approaches, one statistical and one shape-based, for non-rigid motion tracking under the constraints of missing data. The two approaches present a probabilistic framework but are formulated in different ways. Their capacity of tracking the dynamics and estimating the evolution of a curve depends on the observations. In both approaches, the particle filter is integrated with an adaptative dynamical model for quite different objectives. The evaluation of the performance of these approaches depends on the dynamic and the deformation of an object (linear or non-linear). In some particular problems, we have found that one approach gives better results than the other, also depending on the sequences and the deformation undergoes by the curve of the object. This can happen in case of partial deformations like the example of the tennis man, *i.e.* when a part of the curve/shape is only deformed and the other part is rigid: the shape-based approach will give a better estimation of the trajectory than the statistical one. In fact, this difference rests on the formulation of the shape-based approach because it is based on the estimation of the evolution of the remaining points independently, while the other one estimates the deformation of all the shape using its statistical parameters. The main advantage of the shape-based approach is that it takes into account the evolution of all the extrema that represent the essential form of the curve. In the other cases, when all the curves evolve with the same factor of scaling, a global deformation, both approaches are efficient and provide good results (as shown in Section 5).

7 Conclusions

In practice, to improve tracking dynamic contour, learning the motion from a training set is required to define the parameters of the dynamical model. Learning can be handle in the case of plausible motions such as constant velocity or critically damped oscillations. Thus a problem arises in cases of highly non-linear dynamic (e.g. non-periodic oscillation, non-regular acceleration and deceleration, non-constant velocity,...) where we can not define the parameters of the dynamical model. Furthermore, the problem is more complicated in case of no prior knowledge about the dynamic coupled with missing data. For this reason, we have proposed in this paper an adaptative dynamical model which parameters vary with time and are selected from the set of available observations. We also have suggested two approaches to estimate the deformation of an object in video sequences. The results have shown that the objects are successfully tracked over the sequence, despite the highly non linearity of the motion and the constraint of missing data, which proves the robustness of our approaches. Future works will involve extending both approaches to track multiple and varying number of non-rigid objects using a set of multimodal observations.

References

1. Sorenson, H.: Kalman filtering : theory and application. IEEE Press (1985)
2. Jazwinski, A.: Stochastic processes and filtering theory. Academic Press (1974)
3. Doucet, A., Godsill, S., Andrieu, C.: On sequential monte carlo sampling methods for bayesian filtering. *Statistics and computing* (2000) 197–208
4. Doucet, A., Gordon, N., de Freitas, J.: An introduction to sequential monte carlo methods. in *Sequential Monte Carlo Methods in Practice* (2001, Springer-Verlag : New York)
5. Kitagawa, G.: Monte carlo filter and smoother for non-gaussian nonlinear state space models. *Journal of Computational and Graphical Statistics* **1-25** (1996)
6. Blake, A., Isard, M.: *Active Contours*. Springer-Verlag (1998)
7. North, B., Blake, A., Isard, M., Rittscher, J.: Learning and classification of complex dynamics. *IEEE Transactions on pattern analysis and machine intelligence* **22** (September 2000)
8. Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. *Int. J. Computer Vision* (1998)
9. Doucet, A.: On sequential monte carlo methods for bayesian filtering. Technical report, University of Cambridge, UK, Departement of Engineering (1998)
10. Liu, J.S., Chen, R.: Blind deconvolution via sequential imputation. *J Amer. Statist. Assoc* (1995) 567–576
11. Jain, A.: Information and systems science series. *Fundamentals of digital image processing* (1989)
12. Bi, J., Bennett, K.P.: Regression error characteristic curves. *Proceedings of the Twentieth International Conference on Machine Learning (ICML-2003)* (2003, Washington DC)

An Active Contour Model Guided by LBP Distributions

Michalis A. Savelonas¹, Dimitris K. Iakovidis¹,
Dimitris E. Maroulis¹, and Stavros A. Karkanis²

¹ Dept. of Informatics and Telecommunications, University of Athens,
15784, Athens, Greece

`rtsimage@di.uoa.gr`

`http://rtsimage.di.uoa.gr`

² Dept. of Informatics and Computer Technology, Lamia Institute of Technology,

3rd Kilometer, Old National Road, 35100, Lamia, Greece

`sk@teilam.gr`

Abstract. The use of active contours for texture segmentation seems rather attractive in the recent research, indicating that such methodologies may provide more accurate results. In this paper, a novel model for texture segmentation is presented, combining advantages of the active contour approach with texture information acquired by the Local Binary Pattern (LBP) distribution. The proposed LBP scheme has been formulated in order to capture regional information extracted from distributions of LBP values, characterizing a neighborhood around each pixel, instead of using a single LBP value to characterize each pixel. The log-likelihood statistic is employed as a similarity measure between the LBP distributions, resulting to more detailed and accurate segmentation of texture images.

1 Introduction

The automatic segregation of textures within images is generally viewed as an essential first step in various vision applications, such as medical image analysis, industrial monitoring of product quality, content-based image retrieval and remote sensing.

Because of its wide applicability, texture segmentation has been the subject of intensive research in many recent studies [1-5]. However, no known approach is able to consistently and accurately segment textured images [6]. A commonly used strategy for texture segmentation is to extract texture features on a pixel-by-pixel basis and then use some technique to segment the image based on the extracted features and potentially, on some additional spatial constraints. Overall quality of texture segmentation is determined by the quality of both texture features and the segmentation technique.

Early image segmentation approaches have been utilizing boundary-based local filtering techniques such as edge detection operators, which require additional edge-linking operations in order to establish the connectivity of edge segments. This problem has been resolved by employing active contour models [7], which directly result in continuous curves. These models involve the deformation of initial contours towards the boundaries of the image regions to be segmented. A recent active contour model, named Active Contour Without Edges (ACWE) [8] has been gaining increasing

interest due to its advantages: 1) it is region-based, enabling the delineation of regions defined by smooth intensity changes, 2) its level set formulation provides adaptability to topological changes, and 3) it does not impose any significant initialization constraint [8]. However, in the scalar ACWE model the contour evolution depends on the image intensities rather than on the textural content of the image to be segmented. Consequently, the scalar ACWE model cannot discriminate regions of different textures that have equal average intensities.

Latest advances in active contour research focus on using feature vectors to guide contour evolution, as in the case of the extended ACWE model for vector-valued images, proposed by Chan et al [9]. Within a texture segmentation framework, such active contour models use feature vectors that encode the textural content of an image by means of features deriving from Gabor and wavelet transforms [5], [10-11].

The Local Binary Pattern (LBP) distribution, introduced by Ojala et al. [12], offers an alternative approach to spatial texture representation. Unlike the Gabor features, which are calculated from the weighted mean of pixel values over a small neighborhood, the LBP operator considers each pixel in the neighborhood separately, providing even more fine-grained information. In addition, the LBP texture features are invariant to any monotonic change in gray level intensities, resulting in a more robust representation of textures under varying illumination conditions. Comparative studies have demonstrated that the use of LBP distributions may result in higher classification accuracy than Gabor and wavelet features with a smaller computational overhead [12-14].

In this paper we introduce a novel active contour model for texture segmentation guided by LBP distributions. Based on the fact that texture is a local neighborhood property, we have considered using regional information extracted from distributions of LBP values characterizing a neighborhood around each pixel, instead of using a single LBP value to characterize each pixel. In accordance with [15], the similarity between the LBP distributions is estimated by means of the log-likelihood statistic. Moreover, time performance considerations led us to reduce the length of the LBP distributions by limiting the number of pixels participating in the estimation of the LBP values, provided that the resulting LBP operator maintains adequate discriminative capability.

The rest of this paper is organized in five sections. Section 2 briefly reviews the formulation of the LBP operator. The proposed active contour model is presented in Section 3. The results from its application on two-textured images are apposed in Section 4. Finally, in Section 5 the conclusions of this study are summarized.

2 The Local Binary Pattern Operator

We adopt the formulation of the LBP operator defined in [15]. Let T be a texture pattern defined in a local neighborhood of a grey-level texture image as the joint distribution of the gray levels of P ($P > 1$) image pixels:

$$T = t(g_c, g_0, \dots, g_{P-1}) \quad (1)$$

where g_c is the grey-level of the central pixel of the local neighborhood and g_p ($p = 0, \dots, P-1$) represents the gray-level of P equally spaced pixels arranged on a circle of radius R ($R > 0$) that form a circularly symmetric neighbor set.

Much of the information in the original joint gray level distribution (1) about the textural characteristics is conveyed by the joint difference distribution:

$$T \approx t(g_0 - g_c, \dots, g_{P-1} - g_c) \tag{2}$$

This is a highly discriminative texture operator. It records the occurrences of various patterns in the neighborhood of each pixel in a P -dimensional vector.

The signed differences $g_p - g_c$ are not affected by changes in mean luminance; resulting in a joint difference distribution that is invariant against gray-scale shifts. Moreover, invariance with respect to the scaling of the gray-levels is achieved by considering just the signs of the differences instead of their exact values:

$$T \approx t(s(g_0 - g_c), \dots, s(g_{P-1} - g_c)) \tag{3}$$

where

$$s(x) = \begin{cases} 1 & , x \geq 0 \\ 0 & , x < 0 \end{cases} \tag{4}$$

For each sign $s(g_p - g_c)$ a binomial factor 2^p is assigned. Finally, a unique $LBP_{P,R}$ value that characterizes the spatial structure of the local image texture is estimated by:

$$LBP_{P,R} = \sum_{p=0}^{P-1} s(g_p - g_c) 2^p \tag{5}$$

The distribution of the $LBP_{P,R}$ values estimated over an image region comprises a highly discriminative feature vector for texture segmentation [14-17].

3 Active Contour Model Guided by LBP Distributions

The proposed active contour model is inspired by the ACWE model for vector-valued images [9], which uses single point information to guide contour evolution. In what follows, we firstly review this original model and secondly we appose the formulation of the proposed model that uses regional information to guide contour evolution.

3.1 The Original Model

The ACWE model for vector-valued images is based on Mumford-Shah functional [18] and the level set formulation [19]. This model was originally proposed for the segmentation of color images using vectors formed by the RGB values of the pixel intensities [9]. It was later adapted for texture segmentation using Gabor transform coefficients [11]. The model is formulated as follows:

Let u_0 be the original image, defined on a planar domain Ω with real values. Let u_0^i , for $i=1, \dots, b$, be the components that describe the original image u_0 . Let C be the evolving contour. The two averages of u_0^i inside and outside the curve C are denoted as c_+^i and c_-^i for $i=1, 2, \dots, b$. Following [9], an energy functional E is introduced

which, when minimized with respect to $\bar{c}_+ = (c_+^1, \dots, c_+^b)$, $\bar{c}_- = (c_-^1, \dots, c_-^b)$, and C , performs binary segmentation:

$$\begin{aligned}
 E(C, \bar{c}_+, \bar{c}_-) = & \mu \cdot \text{length}(C) + \\
 & \int_{\text{inside}(C)} \frac{1}{b} \sum_{i=1}^b \lambda_i^+ |u_0^i(x, y) - c_+^i|^2 \, dx dy + \\
 & \int_{\text{outside}(C)} \frac{1}{b} \sum_{i=1}^b \lambda_i^- |u_0^i(x, y) - c_-^i|^2 \, dx dy
 \end{aligned} \tag{6}$$

where each value $u_0^i(x, y)$, $i=1, \dots, b$, is defined over a single point (x, y) . For example in [9], $u_0^i(x, y)$ represents the RGB intensities at the point (x, y) , for $i = 1, 2$, and 3 respectively. The positive scalars μ, λ_i^+ and λ_i^- for $i=1, \dots, b$, are weight parameters for each image component. Minimizing the above energy, one tries to segment possible regions in the image with contours given by C and denoted as “inside C ”, from a uniform background denoted as “outside C ”.

In [9] the implementation has been done using the level set method of Osher and Sethian [19], which gives an efficient method for moving curves and surfaces, on a fixed regular grid, allowing for automatic topology changes, such as merging, breaking of curves etc.

The curve C is represented implicitly, via a level set function $\phi(x, y)$ such that $C = \{(x, y) : \phi(x, y) = 0\}$, and $\phi(x, y) > 0$ inside C , $\phi(x, y) < 0$ outside C . The energy E is expressed in level set formulation using the Heaviside function H , which is defined as:

$$H(x) = \begin{cases} 1 & , x \geq 0 \\ 0 & , x < 0 \end{cases} \tag{7}$$

and the Dirac Delta function $\delta(x) = dH(x)/dx$.

$$\begin{aligned}
 E(\bar{c}_+, \bar{c}_-, \phi) = & \mu \cdot \int_{\Omega} \delta(\phi(x, y)) |\nabla \phi(x, y)| \, dx dy + \\
 & \int_{\Omega} \frac{1}{b} \sum_{i=1}^b \lambda_i^+ |u_0^i(x, y) - c_+^i|^2 \, dx dy + \\
 & \int_{\Omega} \frac{1}{b} \sum_{i=1}^b \lambda_i^- |u_0^i(x, y) - c_-^i|^2 \, dx dy
 \end{aligned} \tag{8}$$

Minimizing $E(C, \bar{c}_+, \bar{c}_-)$ with respect to the unknown constant vectors \bar{c}_+ , \bar{c}_- the following relations are obtained, embedded in a time-dependent scheme:

$$\begin{aligned}
 c_+^i(t) = & \frac{\int_{\Omega} u_0^i H(\phi) \, dx dy}{\int_{\Omega} H(\phi) \, dx dy}, \\
 c_-^i(t) = & \frac{\int_{\Omega} u_0^i (1 - H(\phi)) \, dx dy}{\int_{\Omega} (1 - H(\phi)) \, dx dy}
 \end{aligned} \tag{9}$$

i.e. the averages of component u_0^i inside and outside the curve C respectively, for $i=1, 2, \dots, b$ where b is the number of components.

Minimizing $E(C, \bar{c}_+, \bar{c}_-)$ with respect to ϕ , and parameterizing the descent direction by an artificial time, the following Euler-Lagrange equation for ϕ is obtained:

$$\frac{\partial \phi}{\partial t} = \delta(\phi) [\mu \cdot \text{div}(\frac{\nabla \phi}{|\nabla \phi|}) - \frac{1}{b} \cdot \sum_{i=1}^b \lambda_i^+ (u_0^i - c_i^+)^2 + \frac{1}{b} \cdot \sum_{i=1}^b \lambda_i^- (u_0^i - c_i^-)^2] = 0 \tag{10}$$

where a smooth approximation of the Heaviside function H is used, as in [9].

Starting with an initial contour, given by ϕ_0 , at each time step the vector averages \bar{c}_+ , \bar{c}_- are updated and the partial differential equation in ϕ is evolved. More details for the numerical aspects of the level set evolution can be found in [20].

3.2 The Proposed Model

The notion of texture is undefined at single pixel level and it is always associated with some set of pixels [21]. Moreover, as it is stated in Section 2, the single LBP values are texture pattern “signatures” and only their distribution over an image region provides a discriminative feature vector for texture segmentation. This motivated us to formulate the equations of the proposed model using the normalized histogram $N^i(x, y)$, $i=1, \dots, b$, calculated considering regional LBP information, instead of using the single LBP values characterizing each pixel. This regional LBP information is captured by the distribution of the $LBP_{P,R}$ values of all pixels that belong to a $k \times k$ neighborhood centered at the pixel (x, y) . The i -th component, or “bin” of the normalized histogram $N^i(x, y)$, $i=1, \dots, b$ describes the probability of occurrence of a specific texture pattern on each $k \times k$ neighborhood centered at a pixel (x, y) of the considered image region. The total number b of the histogram bins corresponds to the total number of the $LBP_{P,R}$ values and is determined from the number of neighborhood pixels P . It should be noted that in previous vector active contour approaches, the value of each component $u_0^i(x, y)$ of the vector $u_0(x, y)$ is determined from a feature of the single point (x, y) and not from a region feature, as it is the case in the proposed model. For example, in [9] $u_0^i(x, y)$ represents the RGB intensities at the point (x, y) , for $i = 1, 2$, and 3 respectively.

For the sake of efficiency, we choose $LBP_{4,1}$ (Fig. 1) because it involves less complex computations than the standard $LBP_{8,1}$ or other $LBP_{P,R}$ ($P > 8, R \geq 1$) operators and results in a shorter histogram of 16 bins. The $LBP_{4,1}$ operator maintains adequate discriminative capability within the current segmentation framework, as demonstrated by our segmentation results. The use of vector quantization alternatives that have been commonly used instead [16], would introduce a significant computational overhead to the estimation of the feature vectors.

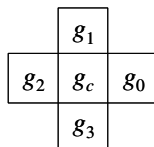


Fig. 1. Local neighborhood of pixels for $LBP_{4,1}$

A rule of thumb suggests that the number of entries for each bin of a histogram should be at least 10. Considering that the $LBP_{4,1}$ produces a 16-bin histogram, the number of entries required for the whole histogram is at least $16 \times 10 = 160$. Therefore $k=13$ corresponds to the minimum neighborhood that satisfies this requirement ($13^2=169 > 160$).

In [15], it is suggested that the similarity between the LBP histograms can be estimated by means of the log-likelihood statistic L . Within our context, the log-likelihood statistic L can be employed as a similarity measure between the LBP normalized histogram $N(x, y)$ and the average LBP histograms \bar{c}_+ and \bar{c}_- of the region inside and outside the contour respectively:

$$L(N, \bar{c}_+) = \sum_{i=1}^b N^i(x, y) \log c_+^i \quad \text{and} \quad L(N, \bar{c}_-) = \sum_{i=1}^b N^i(x, y) \log c_-^i \quad (11)$$

where $N^i(x, y)$ is the i -th bin of the local LBP normalized histogram $N(x, y)$, c_+^i (c_-^i) is the i -th bin of the average LBP histogram \bar{c}_+ (\bar{c}_-), and b is the total number of histogram bins of the considered LBP probability distributions (equal to 16 for the operator $LBP_{4,1}$). As L is an increasing function of similarity of the histograms $N^i(x, y)$ and c_+^i (c_-^i), we use $(1-L)$ as a distance measure between the considered histograms, instead of their squared differences, suggested by equation (6) of the original model. Thus, (6) is replaced by:

$$E(C, \bar{c}_+, \bar{c}_-) = \mu \cdot \text{length}(C) + \int_{\text{inside}(C)} \frac{1}{b} \sum_{i=1}^b \lambda_i^+ (1 - N^i(x, y) \log(c_+^i)) dx dy + \int_{\text{outside}(C)} \frac{1}{b} \sum_{i=1}^b \lambda_i^- (1 - N^i(x, y) \log(c_-^i)) dx dy \quad (12)$$

Minimizing $E(C, \bar{c}_+, \bar{c}_-)$, results in a segmentation of regions characterized by a different average LBP probability distribution than the rest of the image. The positive scalars λ_i^+ and λ_i^- for $i=1, \dots, b$, are weight parameters for the i -th bin of the LBP histograms $N(x, y)$, \bar{c}_+ and \bar{c}_- . Similarly to (6), the regions to be segmented are defined by contours given by C and denoted as “inside C ”, whereas the background region is denoted as “outside C ”.

The Euler-Langrange formulation of (12), which corresponds to equation (10) of the original model becomes:

$$\frac{\partial \phi}{\partial t} = \delta(\phi) \left[\mu \cdot \text{div} \left(\frac{\nabla \phi}{|\nabla \phi|} \right) - \frac{1}{b} \cdot \sum_{i=1}^b \lambda_i^+ (1 - N^i(x, y) \log(c_+^i)) + \frac{1}{b} \cdot \sum_{i=1}^b \lambda_i^- (1 - N^i(x, y) \log(c_-^i)) \right] = 0 \quad (13)$$

where ϕ is the level set function, implicitly representing curve C .

4 Results

The proposed active contour model is implemented and applied for the segmentation of two-texture images, composed of Brodatz textures [22], as well as of natural scenes obtained from VisTex database [23]. In order to evaluate the contribution of the log-likelihood statistic to segmentation accuracy, we perform experiments with: 1) the proposed model employing the log-likelihood statistic, as stated in equation (12), 2) the proposed model employing the squared differences of $N^i(x, y)$ and c^i_+ (c^i_-), as suggested by equation (6) of the original model. Both variations of the proposed model are implemented in Microsoft Visual C++ and executed on a 3.2 GHz Intel Pentium IV workstation. The model constants are generally chosen as follows: $\lambda^i_+ = \lambda^i_- = 750000$, $\mu = 6500$ for the first variation, and $\lambda^i_+ = \lambda^i_- = 750$, $\mu = 6500$ for the second variation. These two sets of values were empirically determined to achieve higher segmentation accuracy in the majority of the two-texture images used. The LBP operator used is $LBP_{4,1}$ and each local LBP histogram is extracted from $k \times k$ neighborhoods with $k=13$, as described in the previous section.

Figures 1-4 illustrate four example results of the application of both variations of the proposed model on two-texture images. The results of the application of the first model variation, employing the log-likelihood statistic, are depicted on Fig. 1(a), 2(a), 3(a), 4(a) whereas the results of the second model variation, employing the squared differences of $N^i(x, y)$ and c^i_+ (c^i_-), are depicted on Fig. 1(b), 2(b), 3(b), 4(b). The

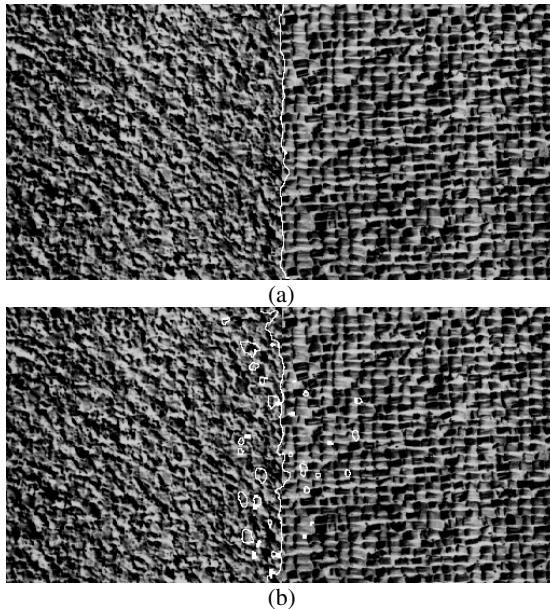


Fig. 1. Segmentation results of the application of the two model variations on the two-texture image D4D84, composed of Brodatz textures [22]: (a) segmentation result of the first model variation, (b) segmentation result of the second model variation

segmentation results obtained by the first model variation, employing the log-likelihood statistic, are very promising. The frames composed of different texture patterns are very well segmented. Moreover, the segmentation quality obtained by the application of the first model variation is generally improved when compared to that obtained by the second model variation, in the cases of Fig. 1,3,4 (in the case of Fig. 2, both variations achieved a practically perfect segmentation result). This improvement indicates that the log-likelihood statistic is more descriptive within the current segmentation framework. The computational cost of our approach varies between 40 and 60 seconds.

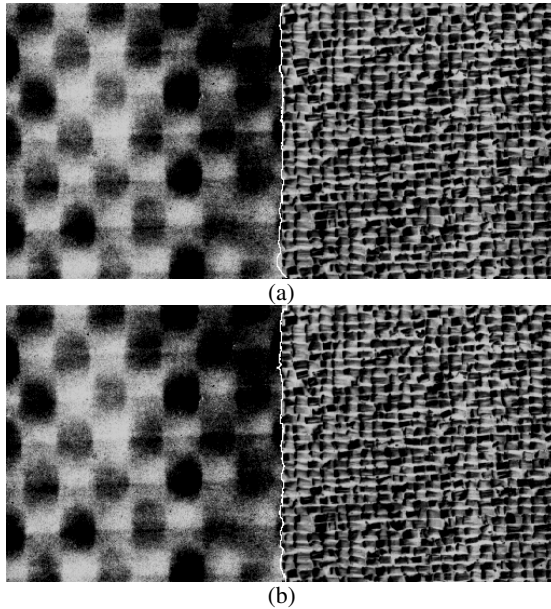


Fig. 2. Segmentation results of the application of the two model variations on the two-texture image D8D84, composed of Brodatz textures [22]: (a) segmentation result of the first model variation, (b) segmentation result of the second model variation

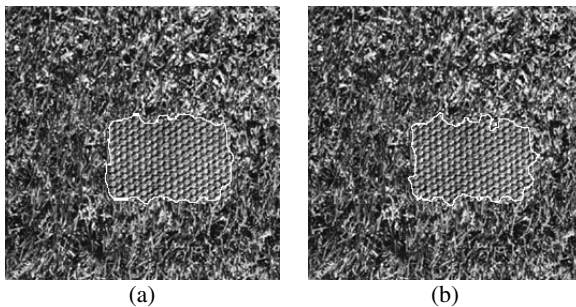


Fig. 3. Segmentation results of the application of the two model variations on the two-texture image D9D77, composed of Brodatz textures [22]. It should be noted that the “ground-truth” shape of the region to be segmented is a rectangular: (a) segmentation result of the first model variation, (b) segmentation result of the second model variation.

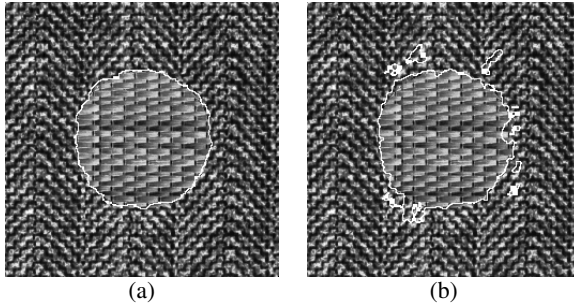


Fig. 4. Segmentation results of the application of the two model variations on the two-texture image D17D55, composed of Brodatz textures [22]: (a) segmentation result of the first model variation, (b) segmentation result of the second model variation

The results illustrated in Fig. 5 show that the proposed active contour model is able to achieve high quality segmentation of natural scenes.

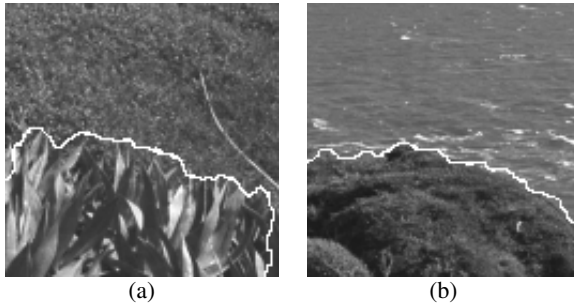


Fig. 5. Segmentation results of the application of the two model variations on natural scenes obtained from VisTex database [23]: (a) GrassPlantsSky.0005, (b) GroundWaterCity.0001

5 Conclusion

In this paper, we presented a novel model for texture segmentation, featuring an active contour approach. The proposed active contour model is guided by the texture information, which is encoded with the use of a local binary pattern scheme. The texture information is extracted from distributions of LBP values, characterizing a neighborhood around each pixel, instead of using a single LBP value to characterize each pixel. As a similarity measure between the LBP distributions, we have used the log-likelihood statistic. We demonstrated that the proposed model achieves high quality segmentation results by applying the model on composite texture images taken from the Brodatz album. Possible future extensions of this work include : 1) an extensive testing on medical images instead of the artificial ones used in this work, 2) the adoption of a quantitative measure for a more accurate evaluation of the segmentation results, 3) test the model performance when adopting the $LBP_{p,R}^{riu2}$ operator introduced

in [15], and 4) extension of the proposed model for the segmentation of multiple-texture images by incorporating the multi-phase ACWE [24].

Acknowledgement

This work was supported by the Greek General Secretariat of Research and Technology and the European Social Fund, through the PENED 2003 program (grant no. 03-ED-662).

References

1. Theodoridis S., Koutroumbas K.: Pattern Recognition, 2nd edn., Academic Press (2003)
2. Mirmehdi M., Petrou M.: Segmentation of Color Textures, IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 2 (2000) 142-159
3. Quing X., Jie Y., Siyi D.: Texture Segmentation using LBP embedded Region Competition, Electronic Letters on Computer Vision and Image Analysis, Vol. 5, No.1 (2005) 41-47
4. Rousson M., Brox T., Deriche R.: Active Unsupervised Texture Segmentation on a Diffusion Based Feature Space, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, Madison, Wisconsin, USA (2003)
5. Sagiv C., Sochen N.A., Zeevi Y.: Integrated Active Contours for Texture Segmentation, IEEE Transactions on Image Processing, Vol. 1, No. 1 (2004) 1-19
6. Clausi D.A., Deng H.: Design-Based Texture Feature Fusion Using Gabor Filters and Co-Occurrence Probabilities, IEEE Transactions on Image Processing, Vol. 14, No. 7 (2005) 925-936
7. Kass M., Witkin A., Terzopoulos D.: Snakes: Active Contour Models, International Journal on Computer Vision, Vol. 1 (1988) 321-331
8. Chan T.F., Vese L.A.: Active Contours Without Edges, IEEE Transactions on Image Processing, Vol. 7 (2001) 266-277
9. Chan T., Sandberg B., Vese L., Active Contours Without Edges for Vector-Valued Images, Journal of Visual Communication and Image Representation, Vol. 11(2002) 130-141
10. Paragios N., Deriche R.: Geodesic Active Contours for Supervised Texture Segmentation, Proceedings of IEEE International Conference on Computer Vision and Pattern Recognition (1999) 2422-2427
11. Aujol J.F., Aubert G., Blanc-Feraud L.: Wavelet-Based Level Set Evolution for Classification of Textured Images, IEEE Transactions on Image Processing, Vol. 12, No. 12 (2003) 1634-1641
12. Ojala T., Pietikäinen M., Harwood D.: A Comparative Study of Texture Measures with Classification based on Feature Distributions, Pattern Recognition, Vol. 29 (1996) 51-59
13. Paclíc P., Duin R., Kempen G.V., Kohlus R.: Supervised Segmentation of Textures in Backscatter Images, Proceedings of IEEE International Conference on Pattern Recognition, Vol. 2 (2002) 490-493
14. Mäenpää, T., Pietikäinen, M.: Classification with color and texture: Jointly or separately?, Pattern Recognition, 37 (8) (2004) 1629-1640
15. 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, Vol. 24, No. 7, (2002) 971-987

16. Pietikäinen M., Ojala T., Nonparametric Texture Analysis with Simple Spatial Operators, Proceedings of 5th International Conference on Quality Control by Artificial Vision, Trois-Rivieres, Canada (1999) 11-16
17. Mäenpää T., Ojala T., Pietikäinen M., Maricor S.: Robust Texture Classification by Subsets of Local Binary Patterns, Proceedings of 15th International Conference on Pattern Recognition, Barcelona, Vol. 3 (2000) 947-950
18. Mumford D., Shah J.: Optimal Approximation by Piecewise Smooth Functions and Associated Variational Problems, Communications in Pure and Applied Mathematics, Vol. 42 (1989) 577-685
19. Osher S., Sethian J.: Fronts Propagating with Curvature- Dependent Speed: Algorithms Based on the Hamilton-Jacobi Formulations, Journal Of Computational Physics, Vol. 79, (1988) 12-49
20. Aubert G., Vese L.: A Variational Method in Image Recovery, SIAM Journal on Numerical Analysis, Vol. 34(5) (1997) 1948-1979
21. Unser, M. , Eden, M.: Nonlinear Operators for Improving Texture Segmentation Based on Features Extracted by Spatial Filtering. IEEE Trans. On Systems, Man and Cybernetics, 20 (4) (1990) 804-815
22. Brodatz P.: Textures: A Photographic Album for Artists and Designers, New York, NY, Dover (1996)
23. Vision Texture Database, MIT Media Lab, www-white.media.mit.edu/vismod/imagery/VisionTexture/vistex.html
24. Vese L.A., Chan T.F.: A Multiphase Level Set Framework for Image Segmentation Using the Mumford and Shah Model, International Journal on Computer Vision, Vol. 50, No. 3 (2002) 271-293

Characterizing the Lacunarity of Objects and Image Sets and Its Use as a Technique for the Analysis of Textural Patterns

Rafael H.C. de Melo, Evelyn de A. Vieira, and Aura Conci

Computer Institute, UFF - Federal Fluminense University,
R Passo da Patria 156, 24210-240 Niteroi, Rio de Janeiro, Brazil
{rmelo, evieira, aconci}@ic.uff.br
<http://www.ic.uff.br/rmelo>

Abstract. An approach is presented for characterize objects and image texture by local lacunarity. This measure makes possible to distinguish sets that have same fractal dimension. In image analysis it can be used as a new feature in the pattern recognition process mainly for identification of natural textures. Illustrating the approach, two types of examples were presented: 3D objects representing approximations of fractal sets and medical images. In the first type, we apply this approach to show its possibility when the objects presents the same fractal dimension. The second type shows that it can be used as a feature on pattern recognition alone in many resolutions.

1 Introduction

One important application of fractals is in the field of image texture analysis. The main aspect of Fractal Geometry used in such application is the concept of fractal dimensions to characterize the texture scaling behavior [3,4,6-10,12]. Like any other measure used to quantify the texture of a region, the obtained results are related with the predicates of the specific quantifier to identify the relevant texture content. For example, the same texture in any linear transformation is easily recognized by measures based on fractal dimensions [3,4]. It is a very important property in a great number of situations and improves the use of this quantify on texture analysis. But this characteristic can be negative depending on the application.

A good use of an identifier is related with the perception of the meaning of what it represents. Fractal dimension represents the complexity of a texture in their description space. However, different textures having same fractal complexity present same fractal dimension [8,10]. Moreover, sets with the same fractal dimension may differ substantially in their structure (Figs. 1 and 2 shows two examples; others can be seen on <http://www.ic.uff.br/~rmelo/projetos.htm>). If the complexities of the textures in analysis are almost the same this characteristic can produce confusion in textural segmentation, then they need additional parameters to be appreciated [2,3]. Although in this case this fractal measure is not adequate, it continues to be the best measure to

identify the same texture under affine transformations. Other fractal based measures can be used to recognize this difference. The proposition here presented to handle this is to describe the set, not only by one fractal characteristic, but by a set of fractal properties with complementary characteristics. It specially discusses how to compute other fractal characteristic: the lacunarity.

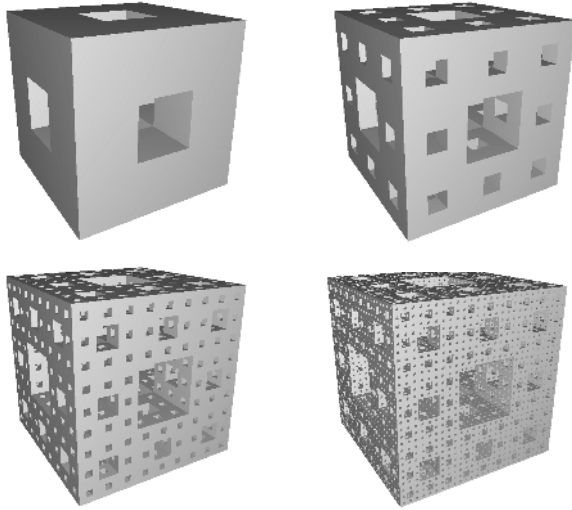


Fig. 1. First steps on reproduction of Sierpinski's sponge. A cube is divided in 27 pieces each one scaled by $1/3$; 7 pieces are eliminated: the middle piece of each face of the cube and the piece on the interior.

1.1 Importance of the Use of More Than One Fractal Measure

It is well known that fractal dimension measures both the irregularity and the fragmentation of sets, which means that fractal properties may be insensitive to topological and affine changes [10]. Moreover, the same fractal dimension, FD, is compatible with very different structures. In other words, objects can exhibit different texture and appearance but still have the same fractal dimension. For example, compare the objects in figure 1 and 2.

Figure 1 represents the first, second, third and fourth initial steps of construction of the fractal named Sierpinski's Sponge [9]. The first step of Sierpinski's Sponge is made dividing each side of a cube by 3, eliminating the central piece of each of the six faces and the interior piece. The others steps are constructed repeating recursively the production. So as 20 pieces are taking and the scale relating each step of the production is $1/3$, its fractal dimension is: $FD = \log 20 / \log 3 \approx 2.7268$.

Figure 2 shows the same steps of a diverse fractal object. The first step of this fractal object is made dividing also each side of a cube by 3, but now the seven eliminated pieces are all connected on the same face as shown in the first image on

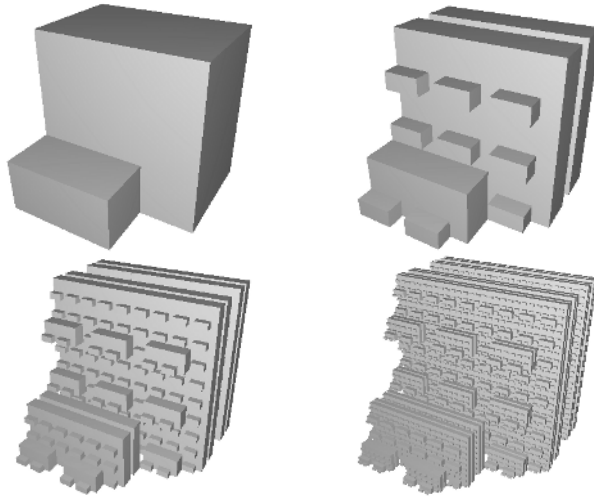


Fig. 2. Initial steps on reproduction of other fractal object made as Fig 1 of 20 parts each one scaled by $1/3$, but grouped without gap. In this object the 7 eliminated pieces are all on the same face where only 2 consecutive pieces kept their positions.

figure 2. The others steps are constructed repeating recursively this production, in a way that now results the objects on the others images on figure 2.

From the construction, the final fractal objects in both figures, 1 and 2, have the same fractal dimension but they are (topologically) completely different. However, this difference is not represented by the FD, it can be characterized in other fractal features.

Empty spaces or gaps and their spatial distribution in more or less regular intervals allowing or not percolation and inter communication are very important aspects. These are manifested in terms of lacunarity and succolarity, respectively. Since lacunarity measures the largeness and regularity of gaps or holes it should be of utility as a feature to differentiate figures 1 and 2. It can also be important in many other textures characterizations and as a feature to be extracted of 3D medical images. In this work, we discuss how it can be used to characterize a texture pattern. An approach to compute it for 2D images or 3D objects is presented. It also considers the efficiency of using a set of local lacunarity (LL) characteristics to texture classification. Examples illustrate practical LL used in real medical images on detection of malignant or benign tumors.

2 Proposed Approach for the Local Lacunarity

Our proposition to compute Local Lacunarity is based on the gliding box algorithm used to analyze the mass distribution on one-dimensional generalized Cantor sets [1]. It considers a box of side s which glides in the object on all possible manners computing the mass distribution with is basic for the lacunarity measure. The notion of position is added to make it possible to distinguish among different part of the same set.

The proposition for compute Local Lacunarity comprehends the following steps:

1 - Objects are first adjusted to an axial aligned bounding box (AABB), example on fig. 3. The size of this box is a function of the object's size: S, in a specific resolution. (If it is an image it is equalized to compensate for possible differences in acquisition conditions before the beginning of local lacunarity, LL, computation. Colour and grey scale information are important aspects on texture [2] and must be considered if available. Lacunarity can be computed on 2D grey-scaled images by threshold or the third coordinate can be used and the images can be seen as collection of voxels. In a specific resolution each voxel can only be considered as empty or full. It is important to stress that the term "local" for lacunarity is related to the AABB position, (i,j), on the original image and to the AABB resolution, r, but mainly to the threshold mass or grey value, t).

2- The gliding box method is used to get the mass probabilities for the AABB in all possible combination of parameters: gliding box edge, s, object resolution, r, and size, S, thresholds levels, t, and position (i,j). Note that the gliding boxes overlap.

3- The probability obtained as a function of all parameters is used to define the lacunarity associated with the local parameters of the object or image.

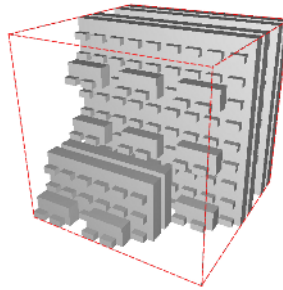


Fig. 3. In dashed, the example of an axial aligned bounding box (AABB) for the third image on figure 2

In the step 2, the incremental analysis of each parameter is associated with the gliding box procedure. In this procedure the first voxel of an $s \times s \times s$ voxels gliding box is initially placed on the corner at first voxel of the images. The computation is performed and then the first voxel of the box is positioned at the second voxel of the images and so on until it reaches the last possible position. Name N this last possible position. The box of $s \times s \times s$ voxels glides over the entire image registering each time the number n_i of non void voxels related to that resolution and its side, s. As s^3 is the maximum possible number of voxels for the box side s, the sequences of the number of non void voxels, $\{n_i\}$, $i \in \{1,2,\dots,N\}$ can be organized to define the frequency of boxes of size s with mass M: $n(M, s)$. For 3D objects, the total number of boxes of size s, $N(s)$, is also a function of the size of the object, S, that is:

$$N(s) = (S - s + 1)^3. \tag{1}$$

The frequency distribution of boxes of size s voxels with mass M , $n(M, s)$, defines a probability function $Q(M,s)$ by dividing it by the total number of boxes:

$$Q(M, s) = n(M, s) / N(s) , \tag{2}$$

where $Q(M,s)$ represents the probability that a gliding box of side s voxels contains M non void voxels, in other words it is easy to show that $Q(M, s)$ satisfies all the axioms of probability [11]. Local Lacunarity (for box side s) is defined by the ratio between the second moment and the first moment square:

$$\Lambda(s) = \sum_{i=1}^N M^2 Q(M, s) / (\sum_{i=1}^N M Q(M, s))^2. \tag{3}$$

When considering all BB position, (i,j) , resolution, r , and possible threshold value, t , it is clear that the above expression is not only a function of the box side s , but related to a set of parameters $(\Lambda_{i,j}(r,t,s))$. Although this method is easily implemented in section 2.1, we use some manual results in order to promptly interpret some aspects, resulting of this measure, that can be useful in many recognition applications. Highest value for a given image will always be found for gliding box equal in size to one voxel, i.e. $s=1$. However, this computation need not be performed since at $s=1$, $Q(1,1)$ represent the occupied ration and $\Lambda(1)$ is the inverse of this value. This value is: (i) only a function of the percentage of occupied sites; (ii) independent of the overall size of the image; and (iii) no related with details of the distribution. Local Lacunarity then must be computed by equations (3) for box side s ranging from 2 to r (or to a representative value).

2.1 Computing the Mass Distribution

Let us compute it for the first object in figure 2, suppose it is represented by $3 \times 3 \times 3$ boxes. The total number of boxes of size $s=2$ that can glide inside the object is 8. If the “gliding” process begin from the most distant position of the viewer to the nearest viewer position, the number of occupied voxels n_i found is: {8, 8, 6, 5, 8, 8, 4, 4}. Then $n(M,s)$ for $s=2$ is defined in table 1.

Table 1. Lacunarity computation for iteration 1 of Figure 2 represented with 9 voxels, considering gliding box with $s=2$

mass: i	frequency $n(M_i,s)$	Probabillity $Q(M_i,s)$	$M_i Q(M_i,s)$	$M_i^2 Q(M_i,s)$
0	0	0	0	0
1	0	0	0	0
2	0	0	0	0
3	0	0	0	0
4	2	0.25	1	4
5	1	0.125	0.625	3.125
6	1	0.125	0.75	0
7	0	0	0	4.5
8	4	0.5	4	32
Σ	8	1	6.375	43.625

Figure 1 - step 1

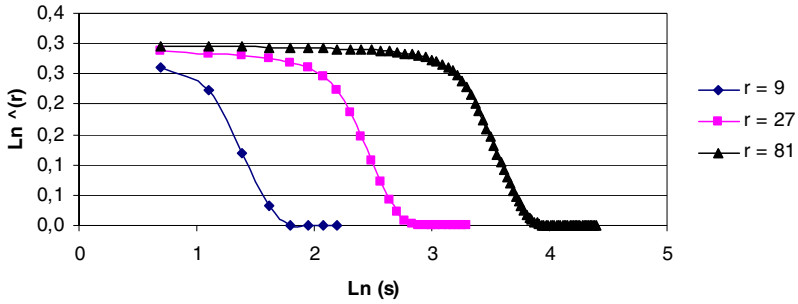


Figure 1 - step 2

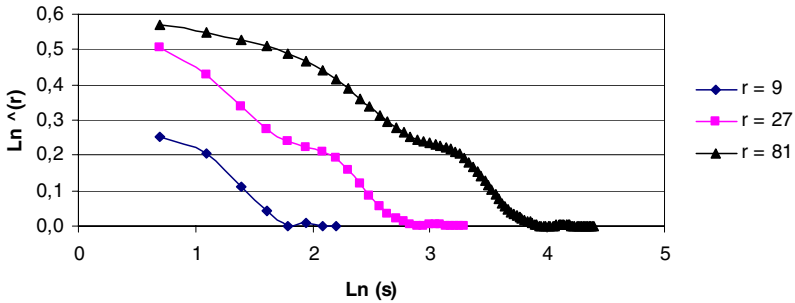


Figure 1 - step 3

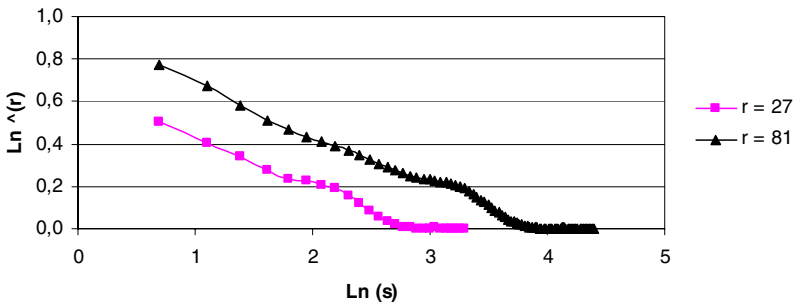


Fig. 4. Log-log plot of lacunarity versus box size for object of figure 1 from the first to the third step. Objects are described at resolution of 9, 27 and 81 voxels.

From table 1 and equation (3) we have: $\Lambda(2) \approx 43.625 / (6.375)^2 \approx 1.073$. The positional parameters in this computation is not important and the threshold level is obviously at void voxel but the obtained result is related with the used AABB resolution of $r=3$. That is, in fact, $\Lambda(3, 0.5, 2) \approx 1.073$.

Figure 2 - step 1

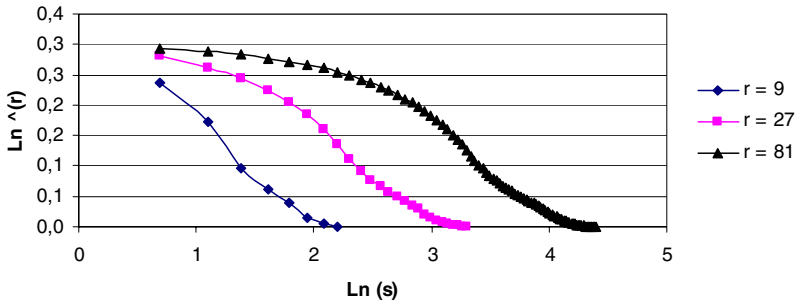


Figure 2 - step 2

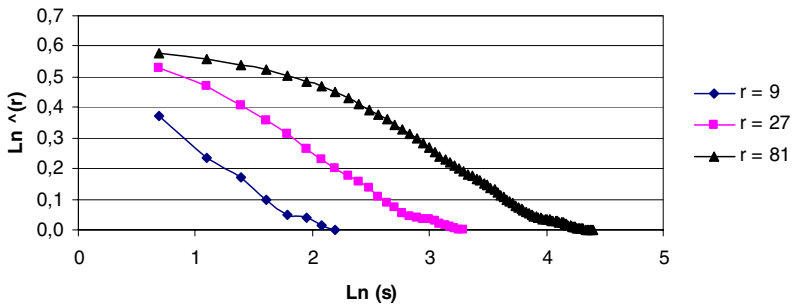


Figure 2 - step 3

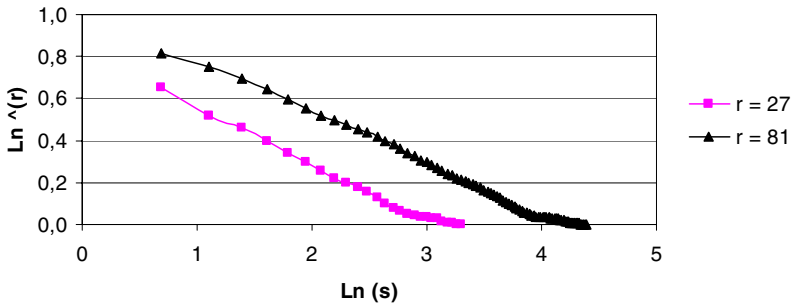


Fig. 5. Log-log plot of lacunarity versus box size for object of figure 2 from the first to the third step when described at resolution of 9, 27 and 81 voxels

Considering the first object in figure 1 and supposing that it is represented by $3 \times 3 \times 3$ voxels, for $s=2$, the number of occupied voxels n_i registered for all position of a gliding box is 4. Then $n(M,2)=8$ and $Q(M,2)=0$ for $M \neq 4$ and $Q(4,2)=1$. So we have: $\Lambda(2) = 1$. That is, for the first iteration of this approximation of Sierpinski's Sponge $\Lambda^1(3,0.5,2)=1$. Another results are $\Lambda^1(81,0.5,2) \approx 1.3449$ and $\Lambda^1(3,0.5,3)=1$ considering both the first iteration on the object construction. For the same object but in other

iteration, as the second, third and fourth object in Figure 1 we have respectively: $\Lambda^2(81, 0.5, 2) \approx 1.7672$, $\Lambda^3(81, 0.5, 2) \approx 2.1619$ and $\Lambda^4(81, 0.5, 2) \approx 2.1577$.

Then, an additional point is that, for real fractals, as these in figures 1 and 2, the LL results change with the iteration. For the same object, but in other iteration, as the second, third and fourth object in these figures we have the values that can be seen on the graph on figure 4 and 5. Note: for real fractal on generation LL is a function of one more parameter, here represented by the index I: $\Lambda_{ij}^I(r,t,s)$.

Figures 4 and 5 shows the results for all possibility of parameters of the objects on figures 1 and 2, that is $\Lambda^I(r, 0.5, s)$ for $r=9,27,81$; $s = 2,3,4, \dots,80,81$ and $I=1,2,3$. The entire object is considered, so the positional parameters (i,j) is irrelevant. It is possible to see on http://www.ic.uff.br/~rmelo/down/results_3D_lacunarity.pdf results for all possibility of parameter variations of the objects on figures 1, 2 and also for other objects.

It is interesting to note that all the objects on figures 1 and 2 will be a real fractal only if the generation goes to infinite. As it is impossible for representation of digital objects or images, they are mathematically only approximations of real fractal objects. Their construction is limited by the voxel limit which is digitally or physically our lower limit of representation.

3 Experiments on Medical Images

For natural acquired objects, the role of the iteration can be represented by the scale used in image capture. Figure 6 represents a set of mammogram images (all database is located on our Web page at <http://www.ic.uff.br/~aconci/mam/frameex1.htm>) with different threshold to binary images. They are used to compute the lacunarity curves showed on the log-log plots of figure 7. The values of threshold were obtained heuristically analyzing the brightness areas of the image and testing some values until became to the optimal values represented on the four binary images of fig. 6.

The idea of consider lacunarity for breast cancer is not new, Einstein et al. [6] measures lacunarity for cytology specimens from Papanicolaou protocol on binary images in a classificatory scheme to benign and malignant diagnoses. Here they are used only to illustrate the proposed methodology of compute Local Lacunarity.

Considering also mammograms Conci et al. [5] used classical image processing techniques to form the five elements feature vector using the 2D shape and contour characterization of the images. LL could be used as another feature in this form of pattern recognition.

The lacunarity plots (Fig. 7) shows important features to note: (1) the lacunarity values reflects de degree gaps; (2) the lacunarity curves is more near of a straight line as more self-similar is the image; and (3) the lacunarity values reveals the aspects of gaps distribution over the entire image, permitting detect presence of hierarchical structures, homogeneity in gaps distribution, random or self-similar behavior.

For objects with the same space occupancy ratio (as those from figures 1 and 2) lacunarity decreases (for same resolution and box size) with the increase of regularity of gaps distribution. For $s=1$ all curves with the same percentage of voxels will present the same value. As a result, all curves for a given number of voxels occupied intercept the same point. A completely regular gap distribution, i.e. a translational invariant image presents lacunarity 1 for any box size larger than the unit size of the repeating pattern.

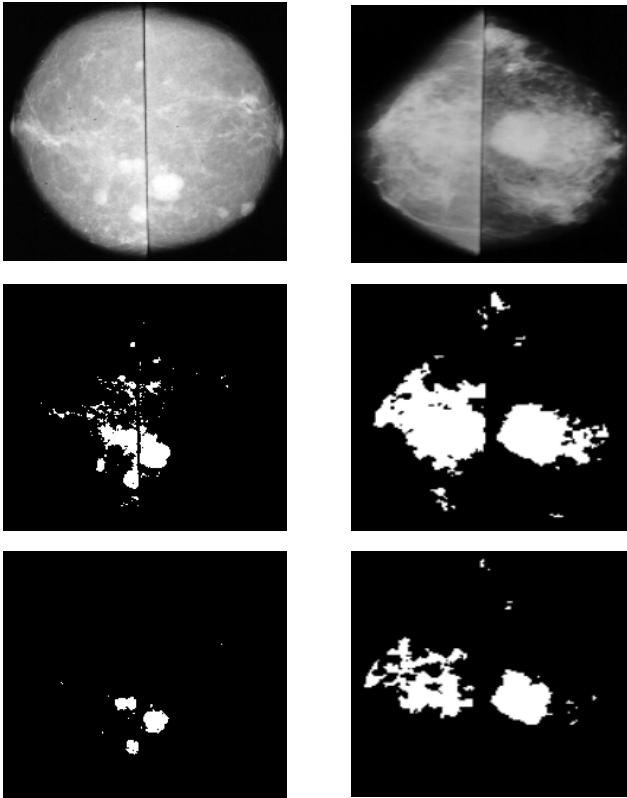


Fig. 6. Two original grey scale image from the mammogram database: Ben3 and Mal13, representing benign (top left) and malignant case (top right). The same images thresholded at 220 and 170 level of the intensity histogram (middle), they are named level 1 on figure 7. Images thresholded at level 2: grey value 240 and 180, respectively.

4 Conclusions

A new idea is presented here: the use of the Fractal Geometry's measure named lacunarity for texture identification in objects or image analysis. Lacunarity has several practical advantages over other indices of texture analysis: method is simple to implement and is not computation expensive compared to other texture calculations; it exhaustively samples the image to quantify changes and self-similarity with scale and it can be used for an analysis of natural or synthetic images.

This work presents also a method for estimation LL of any kind of 3D object or image. It is not a simple extension of the usual lacunarity characterisation of 1D sets because it considers many local aspects as: resolution, generation and box size in images representation. Moreover, image now is an element of the 3D space, which means that its gaps distribution may consider the grey level representing z coordinate.

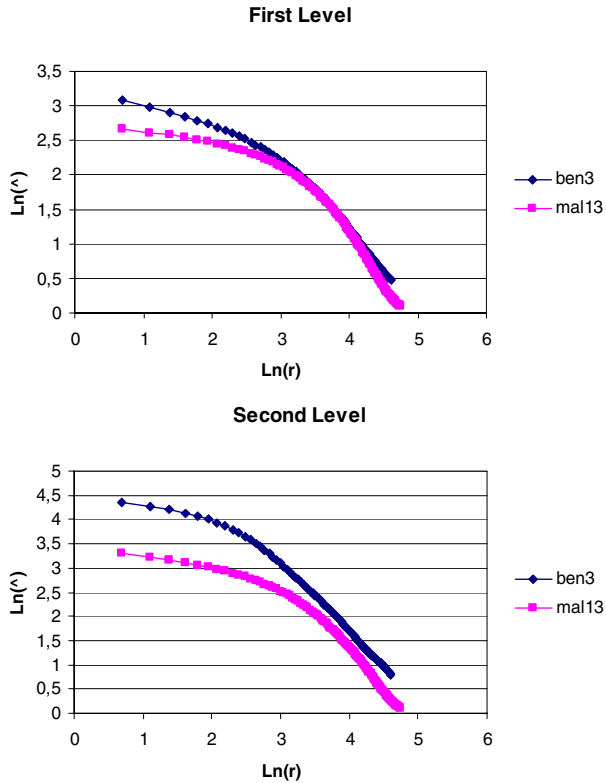


Fig. 7. Lacunarity values for the mammograms in figure 6

The proposed method is used for lacunarity computation of an approximation of fractal objects and real images. The experimental results shows the quality of texture recognition obtained using the lacunarity as a feature extracted from image. Although single values can be used for characterizations aspects of an area like a filter, the use of lacunarity values combined as log-log graphs are much more representative for the characterization of the texture. It is the change of the lacunarity values over different gliding box sizes that yields greater amount of information. These combined values capture pattern over the entire range of scales from the minimum grid to the entire image.

Geometric aspects of an object, a pattern or an image related with fractal Geometry are diverse. Only lacunarity has been considered in this work. Succolarity measures is other aspect of texture related with the sucker, holes or filaments distribution that must be considered in next developments.

This technique can be easily adapted for analysis of 3D images. Results for these and other plots can be seen on the IC Web page at <http://www.ic.uff.br/~rmelo/projetos.htm>. Differently of objects, 2D grey level images presents no gaps unless thresholded. For digital images it is possible identify constrains related to the limits of resolution. These limits are related in the lower bound with the pixel or voxel and in the upper bound with the image resolution, r , in 2D or 3D, that is, the number of pix-

els ($r \times r$) or voxels ($r \times r \times r$) it contains. In other words, in the lower bound lacunarity reflects the degree of space occupancy similarly to the FD. On the opposite side, the larger possible value is the entire image and lacunarity is similarly constrained at this limit.

Considering equation (3), if the denominator (mean), goes to zero, the lacunarity, Λ , goes to ∞ . If the box size contains the entire image ($s = r$), then the variance of the masses in the box is zero and $\Lambda(r)=1$. Although, it is far from the limits that the values measuring the texture present utility, the above discussion of boundary limitation can be useful (e.g. for algorithm verification or as edges representation on graphical visualisation of results). The average mass increases with the box size, then the probability that box masses greatly differ from the average decreases so that relative variance decreases.

The use of a single value of lacunarity estimated considering a single box size is of limited value and meaningless as a basis for comparing different images. The useful feature of lacunarity is the great deal of information gained by computing it over a range of box size. It is especially interesting if resolution, r , also change. For each object, the lacunarity representation can be calculated for box (cubes) sizes ranging from $s=2$ to r incrementing one voxel at the box size each time.

Acknowledgement

This research was supported in part by project CNPq ref. 302649/87-5.

References

1. Allain, M. Cloitre, M., Characterizing the lacunarity of random and deterministic fractal sets, *Physical Review A*, 44 (1991) 3552-3558.
2. Conci, A., Castro, E. M. M., Image mining by content, *Journal of Expert Systems with Applications* 23 (2002) 377-383.
3. Conci A., Nunes, E. O., Multi-bands image analysis using local fractal dimension, *Proceedings of SIBGRAPI-Brazilian Symp. on Comp. Graphics, Image Proc. and Vision* (2001) 91-99.
4. Conci, A., Proença, C. B., A Fractal Image Analysis System for Fabric Inspection Based on a Box-Counting Method, *Computer Networks and ISDN Systems*, 30 (1998) 1887-1895.
5. Conci, A., Soares, L. M., Vianna, A. D. Identification of Benign and Malignant Lesion by Feature Extraction on Mammographic Images, *Applied Mechanics in Americas*, 6 (1999) 53-56.
6. Einstein, J. Hai-Shan Wu, Gil, J.: Self-affinity and lacunarity of chromatin textures in benign and malign breast epithelial cell nuclei, *Physical Review Letters*, 80. (1998) 397-400.
7. Frazer, G. W., Wulder, M. A., Niemann, K. O., Simulation and quantification of the fine-scale spatial pattern and heterogeneity of forest canopy structure: A lacunarity-based method designed for analysis of continuous canopy heights, *Forest Ecology and Management* 214 (2005) 65-90.
8. Keller, J. Crownover, R. Chen, S., Texture description and segmentation through fractal geometry. *Computer Vision Graphics and Image Processing* 45 (1989) 150-160.

9. Mandelbrot, B.: *Fractal Geometry of Nature*, Freeman & Co., San Francisco (1982).
10. Müssigmann, U., *Texture Analysis by Fractal Dimension*. In: J.L.Encarnação, H. O. Peitgen, G. Sakas, G. Englert (eds.): *Fractal geometry and Computer Graphics*. Springer-Verlag, Berlin (1992) 217-230
11. Papoulis, A., *Probability, Random Variables, and Stochastic Processes*. In: *McGraw-Hill Series in Systems Science*, McGraw-Hill, New York (1965)
12. Plotnick, R.E., Gardner, R.H., O'Neill, R.V: Lacunarity indices as measures of landscape texture. *Landscape Ecology* 8 (1993) 201-211.

Adaptive Vision Leveraging Digital Retinas: Extracting Meaningful Segments

Nicolas Burrus and Thierry M. Bernard

ENSTA / UEI

32 Boulevard Victor, 75015 Paris, France

`firstname.lastname@ensta.fr`

Abstract. In general, the less probable an event, the more attention we pay to it. Likewise, considering visual perception, it is interesting to regard important image features as those that most depart from randomness. This statistical approach has recently led to the development of adaptive and parameterless algorithms for image analysis. However, they require computer-intensive statistical measurements. Digital retinas, with their massively parallel and collective computing capabilities, seem adapted to such computational tasks. These principles and opportunities are investigated here through a case study: extracting meaningful segments from an image.

1 Introduction

Designing robust vision algorithms is a serious challenge. The infinite variability of natural images and the difficulty to find precise rules specifying how to solve a -generally trivial for a child- vision problem are greatly contributing to this complexity.

Dynamically adapting algorithms to images they encounter is certainly a way to overcome part of this complexity. Being robust also requires a strict management of algorithm parameters, by relating them to a physical quantity, deducing them from image properties, learning them, selecting them to optimize further treatments through a closed loop process, etc.

Recently, an almost parameterless statistical framework has been proposed [6], relying on the so-called Helmholtz principle, which states that meaningful events are events whose probability to appear in a purely random environment is very low. It seems that human perception follows this rule to some extent, and this framework has been applied notably to gestalts detection with some success [3, 5]. The absence of parameters mainly comes from the fact that no generative probability model of events has to be defined, but only a rarity measure in a well-defined random environment. In addition, some properties of the image can be taken into account to define the random model in which the rarity of other properties will be evaluated, enabling adaptation to the analyzed image.

Such kind of approaches generally requires a considerable number of potentially meaningful events to be evaluated, making real-time processing difficult or impossible. Adaptivity also requires the computation of global quantities, such as

probability distributions over the image, which are time- and power-consuming to obtain using a standard computer. The reason is that pixel data have to be transferred many times, for each pixel, from memory to processor.

To ease such global computations, less standard architectures are needed, that better combine processor and memory, in a more distributed fashion. The latter issue is addressed for more general reasons by the computer architecture community (e.g. [14]). But, in this paper, we focus on artificial retinas (also known as vision chips [11]), which mix processor and memory in an extreme way. Indeed, these are smart imaging sensors, with processing resources in each pixel, thus making massively parallel image array processors without input bottleneck.

On the output side, many retinas are fitted with a global adder (analog as [1] or digital as [10]), able to quickly provide the sum of pixel data over the whole image. The global adder has been used to measure image moments, e.g. for extracting the position of a target. More powerfully, it has been used in a feedback scheme to allow image capture with automatic histogram equalization [12]. We believe that this feedback scheme is worth being systematically extended to image processing : sums provided by a global adder can be surely taken advantage of to better control the way in which images are processed. In particular, it can provide at low cost statistical measures about images in order to make algorithms more adaptive, therefore more robust, as we are looking for. Of course, this only makes sense with programmable retinas, that is retinas with a programmable processor in each pixel - thus allowing versatile image processing - such as the digital retinas we design in our lab [13, 9].

In the present paper, our goal is two-fold:

- show the potential of these general principles through a case study: meaningful segment detection;
- use this experience to improve algorithm/architecture adequation, by motivating future evolutions of both vision system architectures and statistical methods.

In the following, we deal with segment detection in natural images, a standard primitive which can be interesting in artificial environments and which drastically reduces the information contained in an image, while keeping important features. We are looking for an adaptive, parameterless algorithm taking advantage of retina capabilities.

After a global overview of the proposed algorithm in Section 2, Section 3 focuses on segment candidates extraction on digital retinas, then Section 4 gives statistical criteria to decide whether the candidates are meaningful or not. Finally, quantitative results are given in Section 5 and questions raised by this study are discussed in Section 6.

2 Overview of the Algorithm

2.1 What Is a Segment?

Definition 1. *A segment in a cone C is a one-pixel thick connected set of pixels, such that:*

- each pixel has a local gradient direction in C ;
- for each non-extremity pixel, the direction of the vector formed by its two neighbors is also in C .

Gradient vectors are computed using a Sobel operator. To keep cone belonging easy to check, only eight cones are considered, corresponding to the possible angles in a 5x5 discrete neighborhood, as shown in Figure 1. The main steps of the algorithm are as follows:

1. groups of pixels conforming to the definition of segments are extracted as briefly described in Section 3 and their properties (length, mean of pixels gradient magnitude, etc.) are attached to one of their extremities;
2. the extremities are selected by an *a contrario* statistical criterion, as will be detailed in Section 4. The criterion takes into account global image measures and for each segment answers the question: “could a segment with the same properties possibly be observed in a purely random environment?”;
3. segments for which the answer is “no” are reconstructed from their representative extremity, resulting in a binary image of meaningful segments.

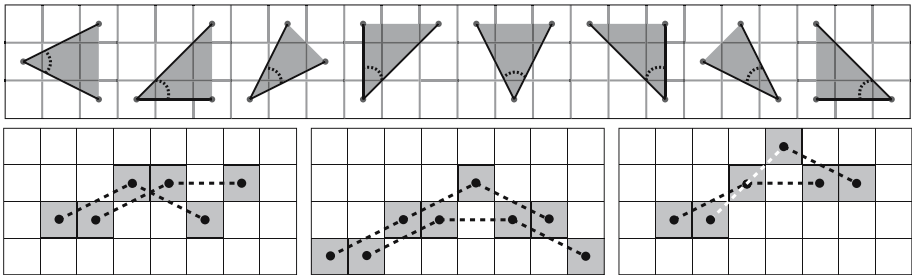


Fig. 1. Top row: the eight overlapping direction cones. Bottom row: illustration of Definition 1 for the first cone (horizontal). Dark pixels must have a local gradient direction in the cone. Dashed lines show the direction induced by the neighbors of each point. Among the three sets of dark pixels, only the left and central ones are segments according to Definition 1, since all dashed lines are within the cone. The right one is not a valid segment since the white dashed line lies outside the cone (too large angle).

3 Candidates Extraction

Segment extraction, the first step of the overall algorithm (see Section 2) is itself performed in three steps, as illustrated in Figure 2:

1. Eight binary images are computed, each representing a direction cone in which segments will be looked for. Any pixel with a determined gradient direction is marked as white in the direction image(s) of which it matches the direction.

2. In each direction image, connected sets of white pixels are made one-pixel thick, such that the remaining pixels lie where the image gradient magnitude is maximal in the orthogonal direction.
3. In each direction image, independently, connected groups of white pixels which match our segment definition are reduced to one of their extremities, to which segments properties are attached. This step is performed by iterative segment erosion: for e.g. horizontal segments, left extremities are removed at each iteration. Before removal, extremities transfer all the information gathered so far to their right neighbor. At the end of this step, extremities support the needed properties of their associated segment.

The extremities are now ready to go through the selection step.

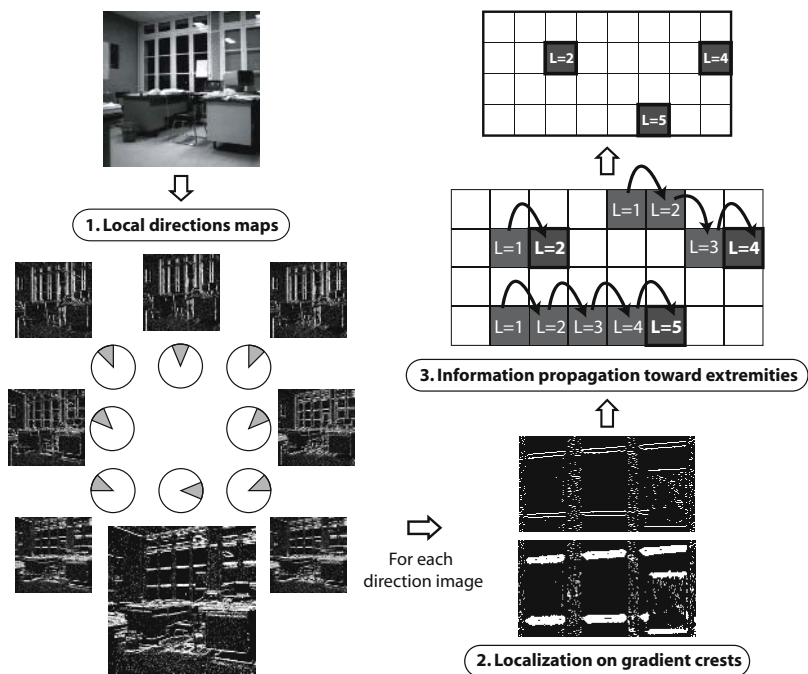


Fig. 2. Overview of the segment extraction algorithm. Step 1 produces eight binary direction images by analyzing local gradient directions in the original image. Each image corresponds to a direction cone and white pixels are pixels whose local direction fits in the cone. Step 2 and 3 are illustrated on portions of the horizontal direction image. Step 2 makes connected sets of white pixels one-pixel thick by only keeping pixels lying where the image gradient magnitude is maximal. Step 3 iteratively propagates segment information (length in the example) from the left extremity to the right one, following the rules of Definition 1. Finally, only right extremities will be fed to the selection step.

4 Candidates Selection

4.1 About the Number of False Alarms (NFA)

The question addressed in this section is: observing a segment with some properties, how to decide whether this segment is meaningful or whether it is just an artefact or coincidence? Two segment properties will be considered, the mean of the gradient magnitude of the segment pixels in Section 4.2 and the length of the segment in Section 4.3. In the spirit of [6], we chose to reason *a contrario*, i.e. instead of computing the probability to observe such a segment in a natural image, we try to answer the question: could the observed segment possibly appear in a noise image? If not, it must be due to a real world phenomenon: object, shadow, etc. To quantify this *a contrario* likelihood, we recall the definition of the number of false alarms of an event.

Definition 2. *The number of false alarms of an event E is the expected number of occurrences of E in a random environment.*

Using the NFA, the notion of ϵ -meaningfulness may be defined, with ϵ a strictly positive (possibly $\ll 1$) real number.

Definition 3. *An ϵ -meaningful event E is an event such as $NFA(E) < \epsilon$. A 1-meaningful event is simply called a meaningful event.*

In practice, choosing $\epsilon = 1$ means that the event is expected to appear less than once in a random context. It is a sound choice as the NFA generally has an exponential behavior w.r.t. event properties, so the dependence on ϵ is rather a log-dependence.

An example is provided in Figure 3 to illustrate how deviations from a random model make events perceptually meaningful.

Finally, to decide whether an event is meaningful in this framework, we need three elements, chosen *a priori*:

1. What kind of events are we looking for?
2. Which event's property should be analyzed?
3. What is the relevant *a contrario* random model?

For segments detection, we have already answered question 1 with Definition 1. Question 2 and 3 will be shortly addressed in Section 4.2 and Section 4.3.

4.2 Selection Based on a Gradient Magnitude Criterion

A natural criterion to start with is based on the contrast between the segment and its neighborhood, an approach comparable to [4]. Whereas [4] was interested in the minimal gradient intensity value along level lines, here we consider the mean gradient value along the segment to be less sensitive to outliers. The gradient magnitude is computed using 2x2 finite differences to avoid creating artificial correlation between pixels (see [2] for detailed explanations).

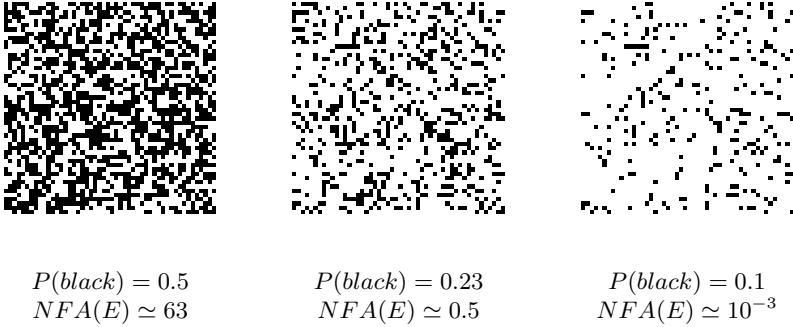


Fig. 3. Illustration of the number of false alarms. Unlike the more complex segments matching Definition 1, the events considered here are simply “perfectly horizontal segments”. The property associated to the event is the length and the *a contrario* random model is an image whose black pixel density is the same as in the original image but where points are spatially independent and uniformly distributed. A segment of 6-pixel length has been artificially added to each image, which otherwise follows the *a contrario* model. As the density of black pixels gets smaller, the number of false alarms of the 6-pixel segment decreases and it becomes an increasingly important deviation from the *a contrario* model. Our perception seems to follow similar rules: the segment becomes detectable with a NFA of 0.5 and obvious for a NFA of 10^{-3} .

The *a contrario* model we choose is a white uniform noise with the same gradient magnitude distribution as in the original image, but where pixels are spatially independent and uniformly distributed. This makes the model adapted to the global gradient properties of the image, while making pixel spatial organization in the original image the source of large deviations. The rarity of a segment will not come from the fact of observing pixels with high gradients in itself, but from the fact that a group of adjacent pixels contains many high gradient values.

Under these assumptions, we can compute a number of false alarms for segments.

Definition 4. Let μ_g and σ_g be respectively the gradient magnitude mean and standard deviation on the whole image. Let $N_{segments}$ be the number of candidate segments detected in the image. Let $\mu(S)$ be the mean gradient value of a segment S and $L(S)$ its number of pixels. Then

$$NFA(S) = N_{segments} \times (1 - \text{normcdf}(\mu(S), \mu_g, \frac{\sigma_g}{\sqrt{L(S)}}))$$

where $\text{normcdf}(x, \mu, \sigma)$ is the normal cumulative distribution function with mean μ and deviation σ applied to x .

This definition comes from the central limit theorem. Under the *a contrario* assumption, a segment can be seen as a collection of $L(S)$ independent and

identically distributed samples of the image, thus the mean of their gradient magnitude should follow a normal law if $L(S)$ is big enough, according to the central limit theorem. Since we have $N_{segments}$ candidates, the expected number of segments having a deviation from the random model at least as large as the one observed for S is the $NFA(S)$ of Definition 4.

We have implemented this selection criterion on a standard computer, but not on digital retinas because of some limitations of the current generation. This is discussed in more details in Section 6. This has led us to consider a different criterion as defined in Section 4.3, enabling a fast implementation on our retina.

4.3 Selection Based on Segment Length

Instead of considering the gradient values along the segment, one might wonder what minimal length is required for a segment to be meaningful, whatever its contrast. The question becomes: in a direction image I_d , how many chains of pixels of length l matching Definition 1 would be expected under an *a contrario* random model?

The choice of the *a contrario* model is somewhat similar to the one of Section 4.2. Taking I_d , the *a contrario* image is an image whose white pixels density is the same as I_d , but where pixels are spatially independent and uniformly distributed. This way, the selection adapts to the global density of pixels sharing the same local directions, and large deviations correspond to large groups of adjacent white pixels.

Unfortunately, even under these fairly simple assumptions, a NFA is analytically difficult to compute. This complexity comes from the rather particular connectivity induced by our definition of segments, which makes the number of candidates difficult to count. Still, we have to find the minimal length above which the NFA will be less than one, depending on the direction image white pixel density. This can be evaluated by stochastic Monte Carlo simulation, that is, by analyzing the actual statistical distribution of the lengths of segments occurring in randomly generated images.

Let I_d be a direction image of size $N \times N$, and p_{white} its white pixel density $\frac{\#whitepixels}{N \times N}$. The following procedure is repeated M times:

1. Generate a random black and white image of size $N \times N$ by drawing independently for each pixel the value white or black according to p_{white} ;
2. Apply on it the segment extraction algorithm of Section 3 ;
3. Store the histogram of the segment lengths.

This results, for one p_{white} value, into a collection of M samples of segment lengths histograms, as depicted in Figure 4. We are looking for the length threshold L_{min} which ensures $NFA(L(S)) < 1$ whenever $L(S) \geq L_{min}$. $NFA(L(S))$ is the expectation of the number of occurrences of segments with length greater than $L(S)$ in random images. It can be estimated from the simulations. Having a $NFA < 1$ means the expected maximal segment length in a random image must be less than L_{min} .

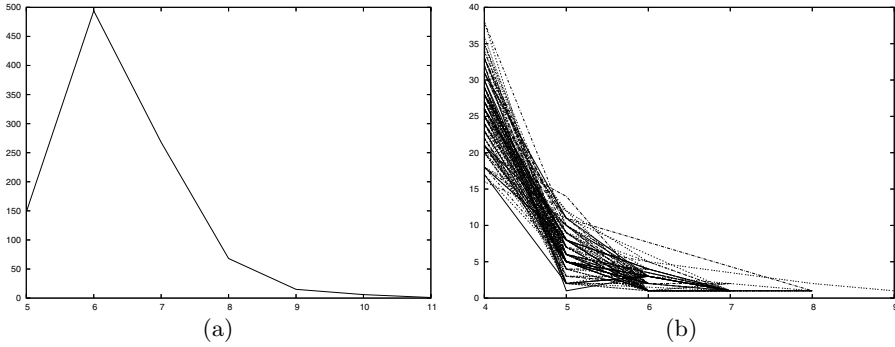


Fig. 4. Horizontal axes are segments length. Vertical axes are the number of occurrences. (a) Histogram of maximal segments lengths for 1000 different uniform noise images with $p_{white} = 0.1$ (b) Histogram of segments lengths for 100 different uniform noise images with $p_{white} = 0.1$, only segments of length greater than four are considered.

From the M simulations above, one can compute a confidence bound on the expected maximal length. Let X_i the maximal length observed in random image i . The empirical mean μ and empirical deviation σ of the maximal length are then:

$$\mu = \frac{1}{M} \sum_{i=1}^M X_i \quad \sigma^2 = \frac{1}{M-1} \sum_{i=1}^M (X_i - \mu)^2$$

Let μ_{true} be the real expectation of the maximal segment length. When M is big enough, the random variable $Y = \frac{(\mu - \mu_{true})\sqrt{M}}{\sigma}$ follows a Student law with $M - 1$ degrees of freedom. We construct a bound on μ_{true} such that:

$$P(Y < t) = \alpha$$

with α the confidence we want. We can get α arbitrarily close to 1 by increasing M and t . Note that α gets exponentially closer to 1 with respect to t , so for $M = 1000$ and $t = 3.1$ the Student law gives $\alpha = 0.999$, leading to:

$$P(\mu_{true} < \mu + 3.1 \frac{\sigma}{\sqrt{1000}}) = 0.999$$

Thus, choosing $L_{min} = \mu + 3.1 \frac{\sigma}{\sqrt{1000}}$ ensures $P(NFA(L(S)) < 1) = 0.999$ whenever $L(S) \geq L_{min}$. Figure 4 shows the typical exponentially decreasing distribution of maximal lengths values.

Finally, running simulations for different p_{white} gives a table of minimal lengths thresholds. Then, the selection algorithm becomes, for each direction image:

1. Estimate p_{white} from the direction image using the global adder;
2. Lookup in a table the corresponding minimal length threshold;
3. Remove extremities associated to segments having a too small length.

5 Quantitative Results

Meaningful segment extraction based on the segment length criterion (see Section 4.3) has been successfully implemented on Pvlsar34, a home-made digital retina of 200x200 pixels, each containing a pixelic Boolean processor with 42 bits of local memory, under SIMD control. To evaluate our algorithm, we bypassed Pvlsar34 capture abilities by transferring standard images into retina memory, scaled to 200x200 pixels and reduced to 64 gray levels to save retina memory. Figure 5 shows an example of segment extraction on an interior scene. Note that there is no free parameter to set in the method, since the segment length thresholds are automatically derived from the direction images densities.

Figure 6 clearly illustrates the benefits of context adaptation. If meaningful segments had been selected on the "house" image with the same threshold as the one derived for the "desk" image, a lot of false alarms would have been obtained, as shown on Figure 6(c).

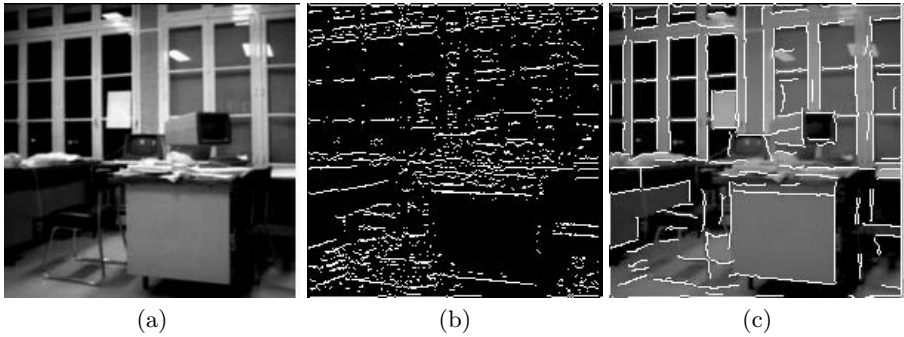


Fig. 5. (a) "Desk" image (b) Horizontal direction image, localized on gradient maxima, $p_{white} = 0.09$ (c) In white: segments which have a meaningful length

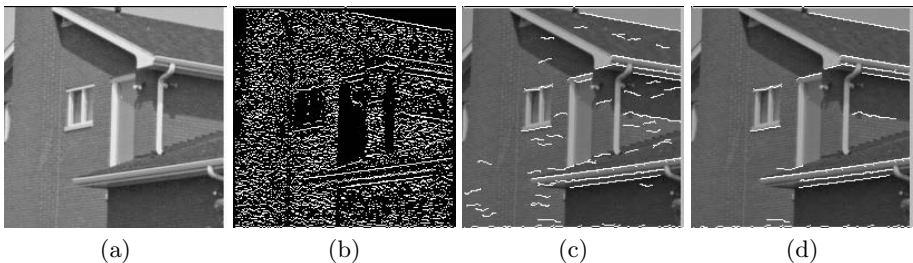


Fig. 6. (a) "House" image (b) Horizontal direction image, localized on gradient maxima, $p_{white} = 0.2$ (c) In white: horizontal segments which would have a meaningful length if the same threshold as for the desk image was used (d) In white: meaningful horizontal segments according to the new threshold

The meaningful segment extraction algorithm runs in real-time on our digital retina Pvlrsar34, at video rate. It runs 10 times slower on an up-to-date personal computer, with 2 images processed per second. This factor of ten seems ridiculously small considering the massive parallelism (40k processors operating together) available in Pvlrsar34. One of the reasons is that Pvlrsar34 is operated at a low frequency of 5MHz, which ensures a very low power consumption of a few tens milliwatts, 3 to 4 orders of magnitude as small as that of a PC! Another reason is examined in the next section.

6 Discussion

6.1 About Statistical Criteria

We try to avoid ad-hoc parameters. But there are still a number of choices which are subject to discussion: the definition of segments, the considered properties, the chosen *a contrario* models, etc. Of course, ideally, those assumptions should be replaced by objective measurements or justifications. What should be noted however is that the nature of the remaining *a priori* is never quantitative but only qualitative. This means the *a priori* decisions always rely on reasoning, never on numerical, empirical values.

Another limitation of current *a contrario* approaches is the global nature of the statistics from which large deviations are measured. There is an underlying assumption of image spatial stationarity, which is obviously not true in the general case. A direct consequence is the so-called blue sky effect, where a very flat zone in the image influences detection in other, shaky, parts. In [2], relying on level lines nesting properties, local meaningful level lines are detected using the statistics of the region associated to their closest surrounding meaningful level line. However, this is not easily applicable in our case.

Finally, we notice that our meaningful segment extraction algorithm does not use so much the retina abilities to compute global statistics though global summations. Using it much more intensively could provide interesting algorithmic and statistical innovations in the future.

6.2 About Candidates and Properties Extraction

Whereas using a digital retina fitted with a fast global adder is a source of algorithmic inspiration, implementing algorithms on it suggests architectural improvements. Here, what are lessons to draw? Whereas gradient and direction-related computations are fast, information propagation toward extremities takes most of the computation time. Indeed, these propagations are done synchronously in Pvlrsar34, and only a few pixels (the extremities) are actually performing useful computations at each iteration. This is clearly under-exploiting massive parallelism. To drastically reduce propagation delays and energy consumption, asynchronous retinas (e.g [8, 7]) have been proposed, allowing efficient computing of regional quantities, which are typical of middle level vision. For

example, the computation of a gradient magnitude mean over a segment would become tractable, thereby enabling more complex properties to be statistically analyzed. More generally, to cope with statistical detection of big groups of pixels, we believe asynchronism will play an important role and we are currently working on the realization of the model described in [9].

6.3 Conclusion

This paper shows a first step towards more adaptive, parameterless and statistically founded algorithms taking advantage of digital retinas philosophy and capabilities. We have developed an original algorithm for the detection of meaningful segments. On the presented images, detected segments indeed seem to be the important ones. These encouraging results are obtained in spite of the relative simplicity of the statistical segment model we have chosen. Implementation on our home-made digital retina has allowed real-time operation but has recalled the limitations of its synchronous SIMD character for middle level vision.

References

1. T. M. Bernard and P. E. Nguyen. Vision through the power supply of the NCP retina. In *SPIE, Charge Coupled Devices and Solid State Sensors V*, volume 2415, pages 159–163, 1995.
2. F. Cao, P. Musé, and F. Sur. Extracting Meaningful Curves from Images. *Journal of Mathematical Imaging and Vision*, 22(2):159–181, 2005.
3. A. Desolneux, L. Moisan, and J.-M. Morel. Meaningful Alignments. *International Journal of Computer Vision*, 40(1):7–23, 2000.
4. A. Desolneux, L. Moisan, and J.-M. Morel. Edge detection by helmholtz principle. *Journal of Mathematical Imaging and Vision*, 14(3):271–284, 2001.
5. A. Desolneux, L. Moisan, and J.-M. Morel. A grouping principle and four applications. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(4):508–513, 2003.
6. A. Desolneux, L. Moisan, and J.-M. Morel. Maximal meaningful events and applications to image analysis. *Annals of Statistics*, 31(6):1822–1851, 2003.
7. B. Ducourthial and A. Mériqot. Parallel asynchronous computation for image analysis. *Proceedings of the IEEE*, 90(7):1218–1229, July 2002.
8. V. Gies, T. M. Bernard, and A. Mériqot. Convergent micro-pipelines: a versatile operator for mixed asynchronous-synchronous computations. In *IEEE International Symposium on Circuits and Systems*, pages 5242–5245, 2005.
9. V. Gies, T. M. Bernard, and A. Mériqot. Asynchronous regional computation capabilities for digital retinas. In *IEEE Workshop on Computer Architecture for Machine Perception and Sensing*, Submitted, 2006.
10. T. Komuro, S. Kagami, and M. Ishikawa. A dynamically reconfigurable SIMD processor for a vision chip. *IEEE Journal Of Solid State Circuits*, 39(1):265–268, 2004.
11. A. Moini. *Vision Chips*. Kluwer Academic Publishers, 2000.

12. Y. Ni, F. Devos, M. Boujrad, and J. H. Guan. Histogram-equalization-based adaptive image sensor for real-time vision. *IEEE Journal Of Solid State Circuits*, 32(7):1027–1036, 1997.
13. F. Paillet, D. Mercier, and T. M. Bernard. Second generation programmable artificial retina. In *IEEE ASIC/SOC Conference*, pages 304–309, 1999.
14. M. B. Taylor et al. Evaluation of the raw microprocessor: An exposed-wire-delay architecture for ILP and streams. In *International Symposium on Computer Architecture*, pages 2–13, 2004.

A Fast Dynamic Border Linking Algorithm for Region Merging

Johan De Bock, Rui Pires, Patrick De Smet, and Wilfried Philips

Dep. TELIN/TW07, Ghent University
Sint-Pietersnieuwstraat 41, B-9000 Ghent, Belgium
`jdebock@telin.UGent.be`

Abstract. In this paper we present our region merging algorithm that is built with special attention on speed but still includes all the necessary functionality to use a wide range of both region based and border based dissimilarity criteria. The algorithm includes a novel method to dynamically link the common borders between two segments during the region merging. The main part of the paper will concentrate on the efficient data structures and functions that are needed to obtain a fast region merging algorithm. Also, all the special situations that can occur in the segment topology are completely covered. We give a detailed report on the execution times of the algorithm and show some of the segmentation results we obtained.

1 Introduction

Image segmentation is the process of partitioning a digital image in meaningful segments, i.e. segments that show a certain degree of homogeneity. This can be homogeneity of any type, such as intensity, color or texture. It is crucial in a number of applications, ranging from image coding and tracking to content-based image retrieval and object recognition. Region merging is one of the most well known and most popular techniques to reduce a set of connected segments (or pixels) to a smaller set by iteratively merging the least dissimilar pair of neighboring segments following a certain criterion that ensures that the final partition consists of meaningful segments. Those criteria can be divided in two classes. The first class consists of region based criteria. They calculate the difference in the statistics of the pixels of two neighboring segments. The second class consists of border based criteria. They incorporate the statistics of the pixels along the common border of two neighboring segments.

In this paper we will present our region merging algorithm that is built with special attention on speed but still includes all the necessary functionality to use a wide range of different criteria from both mentioned classes. Thus the main part of this paper will concentrate on the efficient data structures and functions that are needed to obtain a fast region merging algorithm. Before we start with the explanation of the region merging algorithm we briefly enumerate the steps we took to generate the initial segmentation that is used as the starting set of segments for the region merging algorithm. First we calculate the gradient

magnitude of the input image by convolving it with the first derivative of a Gaussian [1]. Then we regard the gradient magnitude as a topographic landscape and apply the watershed transform on the topographic landscape. These two steps produce the initial segmentation. We used our own fast rainfalling watershed segmentation algorithm [2] to accomplish this task. It has average execution times of approximately 20 ms and worst case execution times of approximately 25 ms (for images of size 512x512 and run on an Intel Pentium 4 2.8 GHz).

In the following sections we will explain in detail how our region merging algorithm is constructed and which decisions we took to make it fast and memory efficient while reasonably generic. The paper is split up in sections following the important modules in the algorithm. In Sect. 2 we will describe and motivate the graph data structure we used to represent the region adjacency graph, the main data structure of a region merging algorithm. In Sect. 3 we will explain how we extracted the inner and outer contour of each segment. In Sect. 4 the extraction of the common border of two neighboring segments out of the previously calculated contours will be discussed. In Sect. 5 the dynamic linking of the borders of two segments that need to be merged will be addressed. In Sect. 6 we end the discussion with the least dissimilar pair of segments calculation and the actual region merging. In Sect. 7 we give a detailed report on the execution times of the region merging algorithm and show some of the segmentation results we obtained. Finally we draw some conclusions in Sect. 8.

2 Region Adjacency Graph Data Structure

Let us first explain what a region adjacency graph is [3]. The region adjacency graph is a planar graph where each node represents a distinct segment of the image partition and where an edge connecting two nodes represents the fact that the two segments associated with those nodes are adjacent (neighbors). We need this graph to get direct access to each neighbor of a segment. The watershed segmentation algorithm produces a segment label image, i.e. an image with for each pixel a label of the segment to which the pixel belongs. So we have to transform the segment label image to a region adjacency graph. First we make sure that every segment is completely connected by horizontal or vertical (four-neighborhood) pixel connection relationships, because later on it will become clear that we get uncontrolled behavior otherwise. We can accomplish this by simply using a watershed algorithm that inherently produces four-neighborhood connected segments, or by transforming the segment label image to that state by splitting up the diagonally (eight-neighborhood) connected segments. Now we construct the graph by doing a raster scan of the segment label image. For every two different neighboring (four-neighborhood) segment labels i, j we encounter, we add the edge (i, j) and the edge (j, i) connecting the nodes i and j to the graph. Consequently the region adjacency graph will be a directed graph. This will be necessary for the dynamic border linking and it also make the handling of the graph data structure easier.

This brings us to the choice we made for the graph data structure. We opted for an array G of linked lists. The edge (i, j) is represented by an element with

value j for the end node field and this element is part of the linked list $G[i]$. Thus each neighbor of a node (segment) i will be in the linked list $G[i]$. This gives us a compact data structure with random access for the start node of an edge and sequential access for the end node of the edge. During the region merging, nodes will only be removed and no new nodes will be created so we do not need a linked list for the start nodes. An adjacency matrix would be an alternative data structure to represent a region adjacency graph, but in this case this would lead to a very sparse matrix with high memory demands. Using a special sparse matrix data structure would lead to exactly the same data structure we just defined.

We use four functions to do the manipulation of the region adjacency graph data structure. The function $AddEdgeC(i, j, w)$ first checks if the edge is not already part of the graph. If it is already part of the graph the function does nothing, otherwise it adds the edge in front of the linked list $G[i]$ with information w . This function is only used when creating the region adjacency graph (during the raster scan). The function $AddEdge(i, j, w)$ just adds the edge without the check, consequently this function is considerably faster than $AddEdgeC$. This function will be used if we are sure the edge isn't already part of the graph. This will always be the case during the region merging because we will first delete an edge, extract the information and then add it again with the updated information. The function $w := DelEdge(i, j)$ looks up the edge, removes it and stores the information in w . The last function $p := FindEdge(i, j)$ looks up the edge and returns a pointer p to the edge.

3 Contour Extraction

To be able to use border based criteria to quantify the dissimilarity between segments, we need to obtain the common border of each pair of neighboring segments. The first step in doing this is the extraction of the inner and outer contour of each segment. To store the border pixels we use linked lists instead of arrays, because we can easily relink those as segments get merged. Copying the contents of arrays would take too much time and random access is not necessary. Before the contour extraction we first calculate several statistics for each segment by doing one raster scan of the input image along with the segment label image and save the statistics in an array. These statistics per segment will be used for the region based criteria. During the same raster scan we also calculate the bounding box of each segment.

We now apply the following steps for each segment to extract the outer contour. We first limit the segment label image to the bounding box of the segment to form a local segment label image. This will reduce the memory requirements for the next steps. Then we enlarge the local segment label image by using pixel replication with a factor three. This will be necessary to cope with segments of only one pixel width during the common border extraction. Now we can easily track the outer contour of the segment by using an eight-neighborhood clockwise contour tracking algorithm. We store all the pixels of this contour in a linked list, each element of the linked list contains the global row and column position and the value of the pixel.

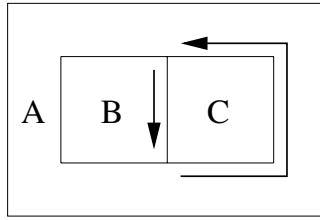


Fig. 1. Special merging case

It is possible that a segment completely surrounds other segments. That segment thus has one or more inner contours. To determine the location of surrounded groups of connected segments, we apply a four-neighborhood connected components algorithm on the local segment label image. For each connected component we then track the inner contour. We have to do this counterclockwise to synchronize them with the outer contours in case one segment of a group of surrounded connected segments would be merged with their surrounding segment. This case is shown in Fig. 1. Here segment *A* and *B* will be merged and they have a common neighbor *C*. The common border of *A* and *C* must follow the same direction as the common border of *B* and *C* to be able to correctly link them, the reason for doing this will be explained in Sect. 5.

4 Common Border Extraction

Now we have to calculate the common border for each neighboring segment pair out of the inner and outer contours and store this information in the region adjacency graph. We will do this directly after the contour extraction of a segment (then we can reuse the local segment label image). We store the common border as a linked list *w.border* in the information *w* of a directed edge. The algorithm works as follows. We scan the linked list of the outer contour of segment *i* pixel per pixel and look within a four-neighborhood for labels of neighboring segments. For each label *j* we find, we add the current pixel at the end of the linked list *w.border* of the corresponding directed edge (i, j) in the region adjacency graph (by using *FindEdge*). We only add the pixel if it is different than the previously added pixel (in the same edge). This is necessary because it is possible to have the same label multiple times as neighbor. We finally repeat all these steps for all the inner contours.

After the previous calculations each directed edge will have at least one pixel in its linked list *w.border*. Otherwise the directed edge would have never been added to the region adjacency graph in the first place. But there are still a few quirks that have to be solved. First we have to make sure that each linked list *w.border* starts with a pixel that is also a start pixel of a border spatially (if the border is not cyclic). We check and solve this by executing the next steps. The start pixel of the linked list must not be eight-neighborhood connected with the end pixel of the linked list. If this is the case we scan the linked list for

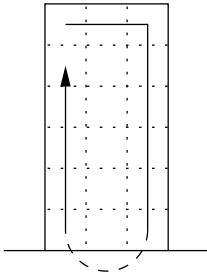


Fig. 2. False connection case

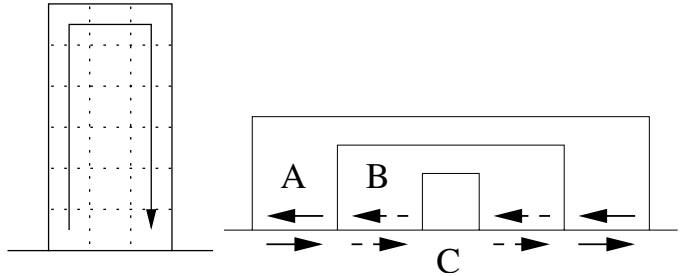


Fig. 3. Separate border parts

two consequent pixels that are not spatially connected. The second pixel then becomes the new start pixel and the first pixel the new end pixel. If we do not find two not spatially connected pixels, then the border is cyclic and then any start and end pixel is good. This method will only work completely if we enlarged the local segment label image by a factor three, otherwise we would get false spatially connected borders in some special cases. In Fig. 2 such a special case is displayed for a segment of 1x2 pixels enlarged by a factor three. The arrow depicts the sequence of pixels in the linked list before and after the start pixel relocation. Correct spatial connection checks would not be possible for smaller enlargement factors.

There is still one special case for a common border we still do not handle correctly. It is possible that a segment has two or more separate connections with the same neighboring segment. The common border thus consists of separate parts. An example of this special case is displayed in Fig. 3, segment *A* and *B* both have separate border parts in common with segment *C*. Consequently we need to split the previously calculated linked lists *w.border* into these separate parts. We store them in *w.border* as an array of linked lists. The search for the pairs of split pixels is also done with the spatial connection check. Without this step it would not be possible to perform correct dynamic linking of these borders.

5 Dynamic Border Linking

When two segments *A* and *B* are merged and when those two segments have a common neighboring segment *C*, we have to link or join the common borders for both sides to form the set of separate border parts corresponding with the new situation. This is shown schematically in Fig. 4. The arrows designate the common borders. Let us assume we have two borders *X* and *Y* that have to be linked, border *X* consisting of *m* separate border parts (linked lists) $X[1] \dots X[m]$ and border *Y* consisting of *n* separate border parts $Y[1] \dots Y[n]$. We first look for the spatial connections between the two sets of border parts. By synchronizing the inner and outer contour earlier, now the end of a border part of *X* can only be spatially connected with the start of a border part of *Y* and vice versa.

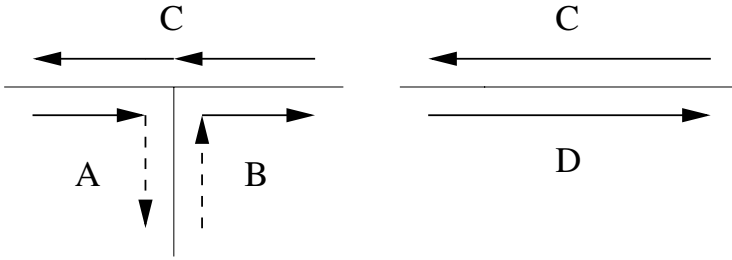


Fig. 4. Dynamic border linking, left: before, right: after

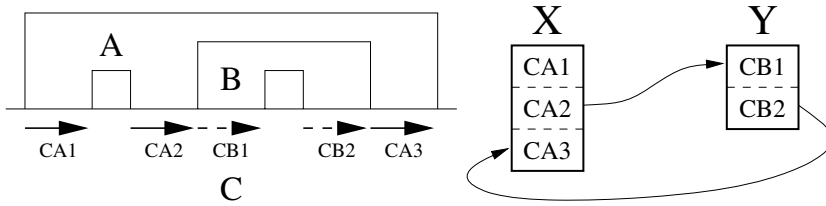


Fig. 5. Schematic view of the dynamic linking process, left: spatial locations of the separate border parts, right: the separate border parts of both borders and the spatial connections (arrows) between them after the search for connections

In the first double loop over the border parts of X and Y we look for the spatial connections from X to Y . In the array of integers $ptrX$ we will store with which border part of Y , a border part of X is connected. In the array of boolean variables $chkY$ we will store which border parts of Y are already connected with their start to a border part of X . So if the end of $X[i]$ is spatially connected with the start $Y[j]$ we set $ptrX[i] := j$ and $chkY[j] := true$. The array $ptrX$ is initialized with -1 and the array $chkY$ is initialized with $false$. Because a border part can only be connected once from X to Y , we can eliminate all the unnecessary spatial connection checks by checking $chkY$ and breaking the inner loop if a connection is found. In the second double loop we look for the spatial connections from Y to X , this is completely analogous. We also count the amount of connections in $conncount$. A schematic view of the dynamic linking process for an exemplary case is displayed in Fig. 5.

After the search for connections, we have all the information we need to link the the border parts to form a new set of border parts. If $conncount = m + n$ then each border part must be linked with another border part, consequently the new border will be cyclic. In this case we just select a start border part and then follow the pointers in $ptrX$ and $ptrY$ to the next border part until we come back to the start border part, meanwhile we link the border parts (linked lists). In any other case we do not have a cyclic new border, and there will be one or more border parts with $chkX[i] = false$ or $chkY[j] = false$. These border parts are the start border parts of the new set of border parts. For each start border

part we follow the pointers in $ptrX$ and $ptrY$ to the next border part until we encounter $ptrX[i] = -1$ or $ptrY[j] = -1$, meanwhile we again link the border parts (linked lists). We store all these new linked lists in an array to form the correct new set of separate border parts.

6 Least Dissimilar Pair Calculation

In this section the least dissimilar segment pair calculation and the actual region merging will be explained. Before we can start with the calculation of the first least dissimilar pair we still have to calculate the initial dissimilarities (one dissimilarity value per criterion) for each pair of neighboring segments. We do this by scanning the complete region adjacency graph and calculating the region based dissimilarities for each edge (pair of neighboring segments). We use the array of segment statistics as input for these calculations. We do the same for the border based dissimilarities, but here we use the statistics directly calculated out of $w.border$. We store all those dissimilarities in the information w of the respective edge.

Now we can start with the calculation of the first least dissimilar pair. We first calculate the minimal edge for each start node, i.e. the edge with the smallest dissimilarity value according to a criterion or a combination of criteria. We will store the values of the minima and the respective start node numbers in an array $minval$ and the respective end nodes numbers of the directed edges in an array $minpos$. We now transform the array $minval$ to an array that satisfies the heap property for the values of the minima. The heap structure is a very efficient data structure to find the next minimum as fast as possible if there are not too many updates in the heap and if the heap is not too small. This concept is also used in [4], but there they used it for the calculation of the minimum of all the edges together. We only use it for the calculation of the minimum of the minimal edges per start node (the array $minval$), because the minimum does not change for most start nodes due to the local nature of the region merging. In fact the only nodes (segments) that will have to be updated after a merge are the two merging segments and their neighbors. For the recalculation of the minimal edge of a start node after a merge we use a normal minimum calculation. Using a heap structure per start node would not be beneficial, because many dissimilarity values will change for the segments in question and because the amount of edges per start node is mostly very limited. It would also use up a considerable amount of additional memory.

The least dissimilar pair of segments is now given by the first element of the heap. The start node number indicates one segment A and the end node number $B := minpos[A]$ indicates the other segment. We now merge the two segments A and B to form segment D (D will take the place of A) by applying the following steps. We first merge the statistics of the segments and update the segment statistics array (in the position of segment A). Then we look for segments that only have segment A and not segment B as neighbor. For all those segments we delete the two directed edges connecting them with segment A , calculate the new

region based dissimilarities, add the edges back to the region adjacency graph for segment D , recalculate the minimum of the start node corresponding with the segment and update the heap with the new minimum. We do not have to update the border based dissimilarities because the common borders remain unchanged for these segments. We repeat the same process for segments that only have segment B and not segment A as neighbor. Next we look for segments that have segment A and segment B as neighbor. For these segments we repeat the same process and additionally we dynamically link the borders and recalculate the border based dissimilarities. To end the region merging iteration we recalculate the minimum of the start node corresponding with segment D (or A) and update the heap with the new minimum. For segment B the heap is updated with ∞ to make sure it disappears. We can repeat this process until a certain amount of segments is reached or until a certain value is reached for the next minimum.

7 Results

To give a good insight in the computational complexity of the different parts of the algorithm, we divided the algorithm in three stages. Stage 1 includes the calculation of the gradient magnitude and the rainfaling watershed segmentation. Stage 2 includes the creation of the region adjacency graph, the calculation of the segments statistics, the extraction of the inner and outer contour, filling the graph with the common borders and filling the graph with the initial region based and border based dissimilarities. Stage 3 includes all the region merging iterations. We tested the algorithm on the well-known test image Peppers 512x512 and measured the execution times of the three different stages. The amount of segments in the initial segmentation (used as input for the region merging and produced by stage 1) is the most influential factor in the total execution time. Therefore we tuned the parameters of stage 1 to get more or less a certain amount of segments. We then logarithmically lowered the amount with a constant factor two. The region merging was stopped at 20 remaining segments and the criterion used for the region merging was the minimal mean squared error increase criterion [4, 5]. All the execution times of this test procedure are shown in Table 1. The complete algorithm is implemented in C and this test was run on an Intel Pentium 4 2.8 GHz. Our algorithm calculates several region

Table 1. Execution times for Peppers 512x512

Number of segments after stage 1	Execution times (in seconds)			
	Stage 1	Stage 2	Stage 3	Total
24014	0.07	1.27	1.98	3.32
12007	0.07	0.87	1.14	2.08
6001	0.10	0.62	0.46	1.18
3008	0.12	0.46	0.20	0.78
1502	0.17	0.35	0.10	0.62
750	0.17	0.29	0.05	0.51

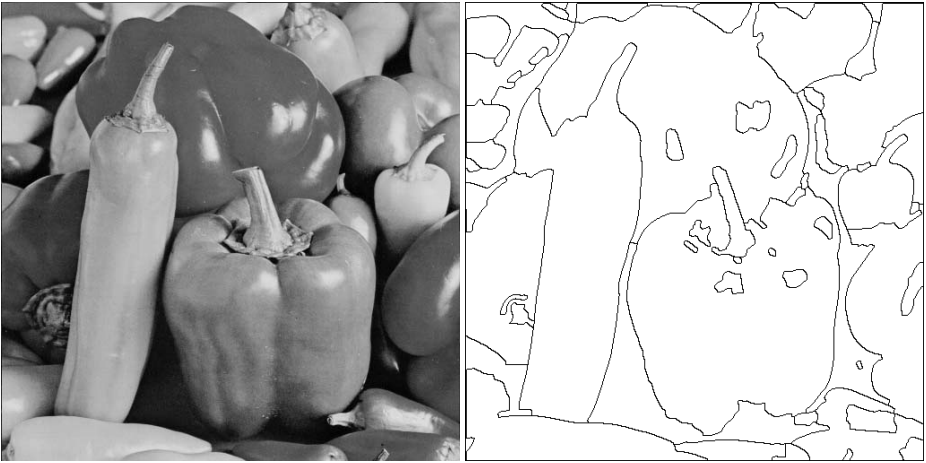


Fig. 6. Left: Peppers 512x512, right: the segmentation produced by our region merging algorithm

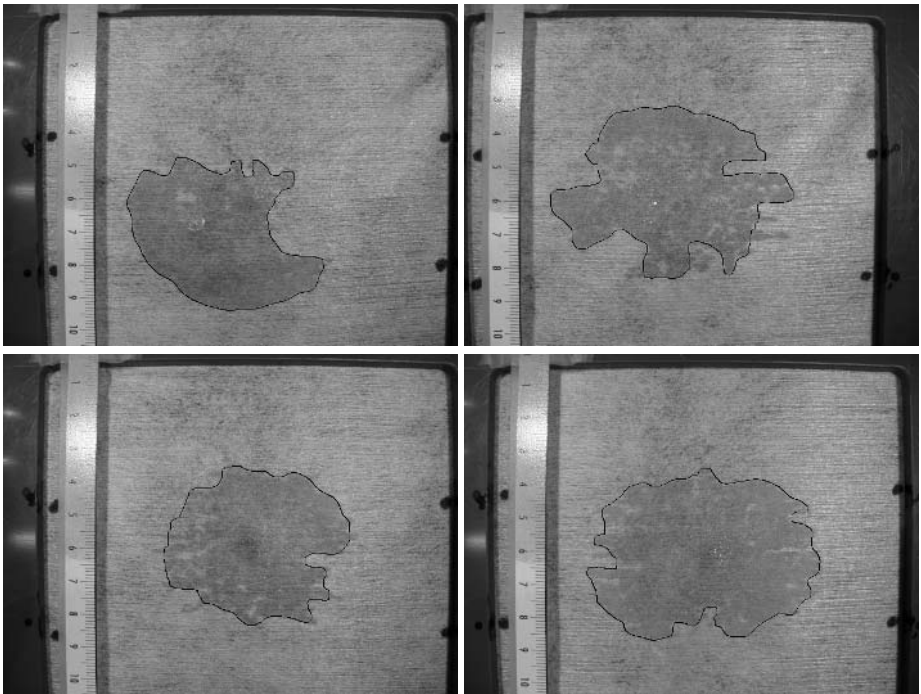


Fig. 7. Segmentation of a spot on a non-woven textile

based and border based criteria, including the already mentioned minimal mean squared error increase criterion, the Fischer distance [6], the average gradient on the common border [5], ... , also any combination of the criteria can be used

as a criterion. A segmentation result on Peppers 512x512 is displayed Fig. 6, here we used the minimal mean squared error increase criterion combined with the minimum of the gradient on the common border. We also included marker propagation in our region merging algorithm. This gives us the possibility to use marker images to guide the segmentation or to perform region of interest segmentation. We used the region of interest segmentation to segment the spot in the following application. The goal of this application was to quantify the hydrophile character of a non-woven textile by measuring the area of the spot left by the water. Some of the segmentation results we obtained are displayed in Fig. 7. Here we used a border based criterion that measures the median of the gradient in a window along the common border to get extra noise robustness without losing too much detail. This criterion exploits the fact that we know the order of the border pixels at any time.

8 Conclusion

We have developed a fast region merging algorithm that includes all the necessary functionality to use a wide range of both region based and border based dissimilarity criteria. It uses a novel method to dynamically link the common borders between two segments during the region merging. We showed that the algorithm correctly handles all the special cases that can occur in the segment topology. We gave a detailed report on the execution times of the algorithm and showed some of the segmentation results we obtained. Future work could be the optimization of the contour extraction algorithm and the inclusion of more criteria and input features.

References

1. Canny, J.: A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **8** (1986) 679–698
2. De Bock, J., De Smet, P., Philips, W.: A fast sequential rainfalling watershed segmentation algorithm. In: *Advanced Concepts for Intelligent Vision Systems, 7th international conference, Antwerp, Belgium*. Volume 3708 of *Lecture notes in computer science.*, Springer (2005) 477–484
3. Pavlidis, T.: *Structural Pattern Recognition*. Springer (1977)
4. Haris, K., Efstratiadis, S., Maglaveras, N., Katsaggelos, A.: Hybrid image segmentation using watersheds and fast region merging. *IEEE Transactions on Image Processing* **7** (1998) 1684–1699
5. Pires, R., De Smet, P., Philips, W.: Watershed segmentation and region merging. In: *IS&T/SPIE Electronic Imaging, Visual Communications and Image Processing*. Volume 5308. (2004) 1127–1135
6. Tremeau, A., Colantoni, P.: Regions adjacency graph applied to color image segmentation. *IEEE Transactions on Image Processing* **9** (2000) 735–744

Fast Sub-pixel Motion Estimation for H.264

Hong Yin Lim and Ashraf A. Kassim

National University of Singapore, 4 Engineering Drive 3 Singapore 117576
ashraf@nus.edu.sg

Abstract. Due to the variable block sizes and multi-reference frame used in the H.264/AVC standard; the motion estimation process becomes even more computationally intensive, resulting in a very low encoding speed. To overcome this low encoding speed, fast motion estimation algorithms such as UMHexagonS [1] and EPZS [2] have been proposed. Since the integer-pixel motion estimation speed has significantly decreased, the fractional or sub-pixel motion estimation speed is no longer non-negligible. We propose a fast sub-pixel motion estimation algorithm using an adaptive rood pattern based on the fractional motion vector of adjacent blocks and also a simplified small diamond search. Our algorithm is able to reduce the number of sub-pixel search points by more than 50%, while restricting the PSNR loss to less than 0.1 dB, compared to the hierarchical fractional pixel search.

1 Introduction

The sub-pixel motion estimation (SPME) is used to improve the accuracy of the block matching method, where each sub-pixel interpolates its value from the original integer pixels. In H.264/AVC, the final motion vector (MV) has quarter-pixel accuracy by default. This allows the final MV to have a more precise range of: $(MV-0.75) \leq MV \leq (MV+0.75)$, with a ± 0.25 step-size in value. By using SPME, the reconstructed video has a 1-2 dB improvement in quality.

In the current SPME algorithm (known as *hierarchical fractional pixel search* (HFPS) [1]) provided in the JM reference software, the 8 half-pixel positions around the integer pixel MV are initially evaluated. The sub-pixel center position- (0,0), which corresponds to the integer MV value is also reevaluated if Hadamard transform [3] and adaptive block-size transforms (ABT) [3] are used. The best MV of the half-pixel position is then selected. The search then shifts to the best half-pixel MV, which acts as the search center for refining the MV to quarter-pixel accuracy. The 8 quarter-pixel surrounding the half-pixel MV is evaluated and the best MV, which has quarter-pixel accuracy, is obtained. These steps are illustrated in Fig. 1.

In the HFPS method, 8 interpolations using a 6-tap filter and 8 linear interpolations are performed to obtain the half-pixel values and quarter-pixel values respectively. In addition, the 16 sub-pixel points are evaluated using the *Sum of Absolute Difference* (SAD) operations. The computational complexity is further

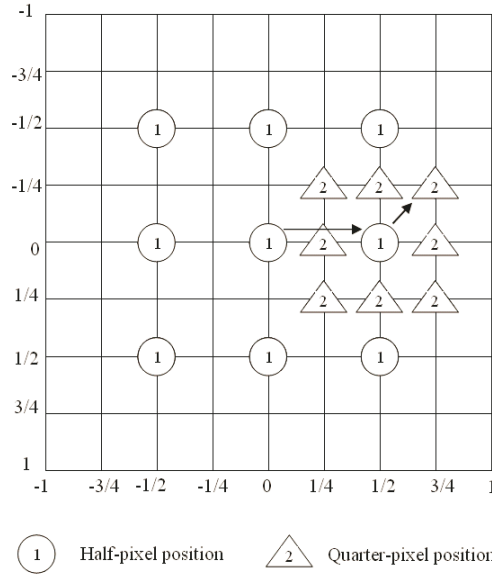


Fig. 1. Hierarchical fractional pixel search (HFPS)

increased if Hadamard transform and ABT is used. In this case, the SPME becomes more significant in the motion estimation process with the improvement in the integer motion estimation speed.

A Center-biased Fractional Pel Search (CBFPS) algorithm [1] has been proposed to reduce the computational complexity of HFPS. In CBFPS, the sub-pixel search is treated as a normal motion search with a maximum search range of 3 and the distance between the points to be in quarter-pixel (± 0.25) unit. The sub-pixel search begins by evaluating the predicted fractional pel MV and the integer-pel MV, which corresponds to the (0,0) position in the sub-pixel search window (see Fig. 1). The predicted fractional pel MV is obtained using:

$$frac_pred_mv = (pred_mv - mv) \% \beta \tag{1}$$

where $pred_mv$ is the basic predicted MV used in H.264/AVC and is obtained as the median value of the MV of the left, top, top-right adjacent blocks. mv is the integer pel MV, $\%$ is the remainder operation and β is 4 for quarter-pel accuracy.

The CBFPS algorithm is illustrated using Fig. 2. In CBFPS, the best predicted sub-pixel point from the predicted fractional MV and the (0,0) forms the search center where further refinement search is performed using the moving Small Diamond Search pattern (SDSP) [4] method. In the moving SDSP method, the small diamond pattern (Fig. 3) is placed on the search center and its points are evaluated. The best point obtained forms the new search center for further search using the small diamond pattern. This is iteratively done until the minimum cost function [3] is located at the center of the pattern.

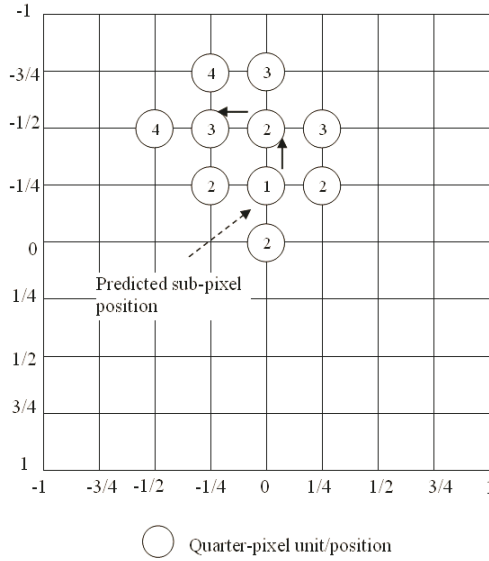


Fig. 2. Center-biased Fractional Pel Search (CBFPS)

In this paper, we propose an extension to the CBFPS method, which have similar quality to the CBFPS, but reduce the number of sub-pixel search points. Our algorithm uses an adaptive rood pattern and also simplifies the moving SDSP method.

2 Adaptive Rood Pattern

The ARPS algorithm [5] introduces an adaptive rood pattern (Fig. 4), where the size of the rood pattern is adaptively calculated based on the motion activity of the adjacent blocks: left, top, top-left, top-right. The rood arm size is calculated based on the maximum distance between the MVs of the adjacent blocks, since it can effectively represent the dynamic range of the local motion movement. The following equation is used to calculate the rood arms size:

$$R_x = \frac{1}{2} (Max [MV_x] - Min [MV_x]); R_y = \frac{1}{2} (Max [MV_y] - Min [MV_y]) \tag{2}$$

where MV_x and MV_y are the horizontal and vertical components respectively of the adjacent blocks' MVs. The result from (2) is rounded-up to the nearest integer value. For example if the calculation returns a value of 0.5, the respective rood arm size is equals to 1.

Our fast SPME method uses the adaptive rood pattern to refine the fractional motion search around the best predicted sub-pixel position, before using the SDSP. However in the calculation of the rood arm size, we only use the MVs that has a significantly close value to the current integer-pel MV. In our study,

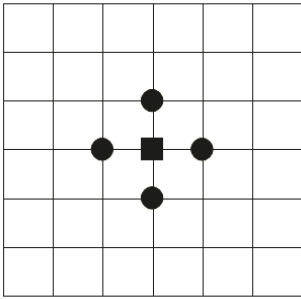


Fig. 3. Small Diamond pattern

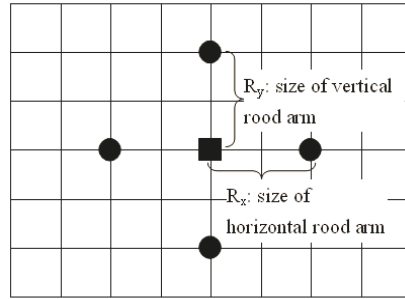


Fig. 4. Adaptive road pattern

we define these significantly close MV as integer-pel MVs that has an absolute value of less than 3 from the current MV value.

Select MV if $(\|MV(x)_{cur} - MV(x)_{adj}\| < 3 \text{ AND } \|MV(y)_{cur} - MV(y)_{adj}\| < 3)$

where $MV(x)_{cur}$, $MV(y)_{cur}$ and $MV(x)_{adj}$, $MV(y)_{adj}$ are the horizontal and vertical components of the current integer-pel MV and adjacent MV respectively.

Consider the following case: the MV of one of the adjacent blocks, MV_1 is = 10.25 (0.25 is the fractional-pel MV), while the current integer-pel MV is = 0. Therefore, we would expect that the fractional-pel unit of MV_1 will not be highly correlated with the current fractional motion due to the significantly different integer-pel MV values.

A study is done to evaluate the effectiveness of the use of the adaptive road pattern for the fractional motion search. A comparison is performed between the following methods:

1. Method 1: The use of only the predicted fractional MV and integer-pel MV, (0,0) in sub-pixel search window without any patterns.
2. Method 2: The use of the adaptive road pattern on the best position from Method 1.

The best fractional MV obtained from the methods is compared to the optimal fractional MV of HFPS. Figures 6-8 shows the percentage of the correct fractional MV, which is defined as the percentage of the best fractional MV having the same value as the optimal fractional MV, while Table 1 shows the average sub-pixel search points evaluated and the entropy of the difference between the best fractional MV and the optimal fractional MV. The results shown in Figures 6-8 and Table 1 is the average of the results encoded using quantization parameter (QP) of 28, 32, 36, 40 and each sequences is encoded using a frame-rate of 30 fps.

Figures 6-8 and Table 1 shows that with the use of the adaptive road pattern, the percentage of the correct fractional MV improves by 1-10%, while the entropy difference between the best fractional MV obtained and the optimal fractional MV is reduced by 3-13%. As seen in the results for the high-motion

sequences such as *Tempete* and *Mobile*, the increase in the percentage of the correct fractional MV by using the adaptive rood pattern is quite significant, with an improvement of 10%. At the same time, the increase in the number of search points from using the rood pattern is not too high, an addition of < 1.7 points. Therefore, it is advantageous to use the adaptive rood pattern in improving the fractional motion search.

3 Simplified Small Diamond Search

The use of the rood pattern in Sect. 2 has improve the accuracy of the fractional motion search, as indicated by the increase in the correct fractional MV obtained and the decrease in the entropy difference. Hence, it is very probable that the best fractional MV obtained is close to the optimal fractional MV.

In this case, we propose that the small diamond pattern (SDP) is only used once for the refinement search. The SDP is placed onto the best sub-pixel position obtained from the rood pattern search and its points are evaluated. If the minimum cost function is located at the center of the pattern, the search is terminated. Otherwise, 2 extra corner points corresponding to the edge with the minimum cost function is evaluated. This is illustrated in Fig. 5.

Similar to the study performed for Sect. 2, we also evaluate the effectiveness of the proposed simplified small diamond search for the sub-pixel motion estimation. The comparison done is performed between the following methods on the use of different refinement patterns:

1. Method 3: Simplified Small Diamond Search.
2. Method 4: Moving Small Diamond Search pattern.

The result of this study is shown in Figs. 6-8 and Table 1. As seen from Figs. 6-8, the use of the proposed simplified Small Diamond Search is able to significantly improve the overall percentage of the correct fractional MV by 7-25%. This is especially prominent for the high-motion sequences, where the improvement of the correct fractional MV is $\geq 20\%$. The entropy difference is similarly decreased by a significant amount of 30-60%. This improvement comes at the modest expense of an increase of fractional search points by only 3.7-4.2 points.

By using the more computational intensive moving SDSP, the increase in percentage of correct fractional MV, compared to our proposed simplified Small Diamond Search is 1-10%, while the entropy difference is further decreased by 20-40%. However, this comes at the expense of an increase in fractional search points by 0.5-1.7 points. Therefore, if computational speed is not of the main concern, then the use of the moving SDSP can be considered.

From Figs 6-8, it is observed that the use of our proposed simplified Small Diamond Search already returns quite a satisfactory result. In most cases, the number of correct fractional MV obtained is $> 75\%$. In fact for the low-motion sequences, the number of correct fractional MV obtained is $\geq 94\%$. This shows that the use

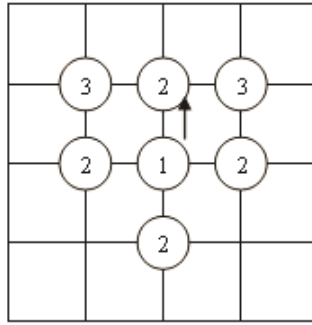


Fig. 5. Simplified Small Diamond Search for sub-pixel motion estimation

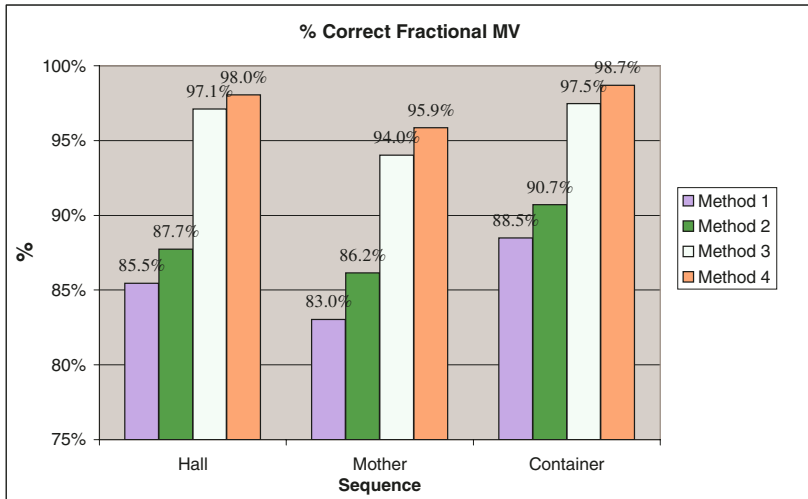


Fig. 6. Percentage of correct fractional MV obtained from different implementations for low-motion sequences

of the simplified Small Diamond Search is already sufficient to obtain good results, and the advantage of using the moving SDSP is only marginal. Therefore, we firmly propose the simplified Small Diamond Search for use in the SPME.

4 Results

Our experiments were carried out using the H.264/AVC reference software [6], where the H.264/AVC main profile is used, and the following encoding parameters are switched on: Hadamard transform, RD Optimization and CABAC encoding. We use the UMHexagonS algorithm for the integer-pel motion estimation and the default quarter-pel accuracy is used for the fractional motion search. The results

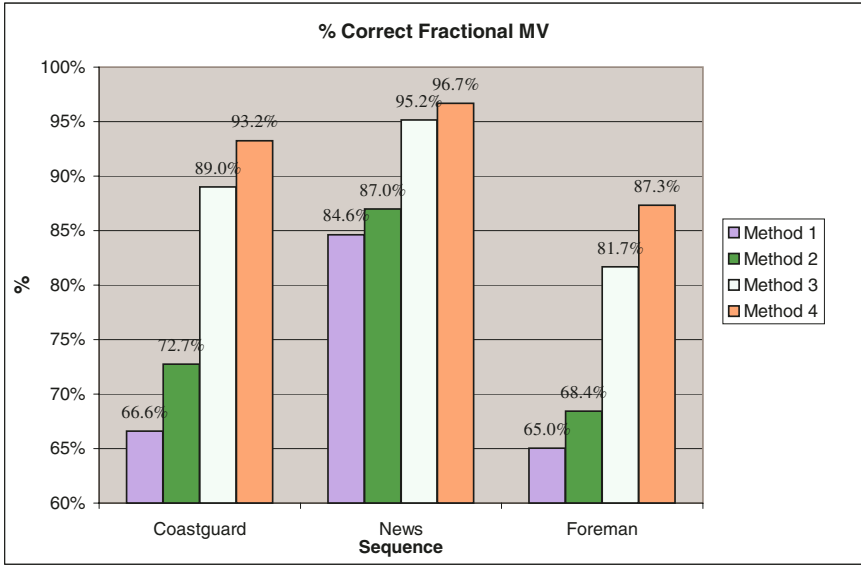


Fig. 7. Percentage of correct fractional MV obtained from different implementations for medium-motion sequences

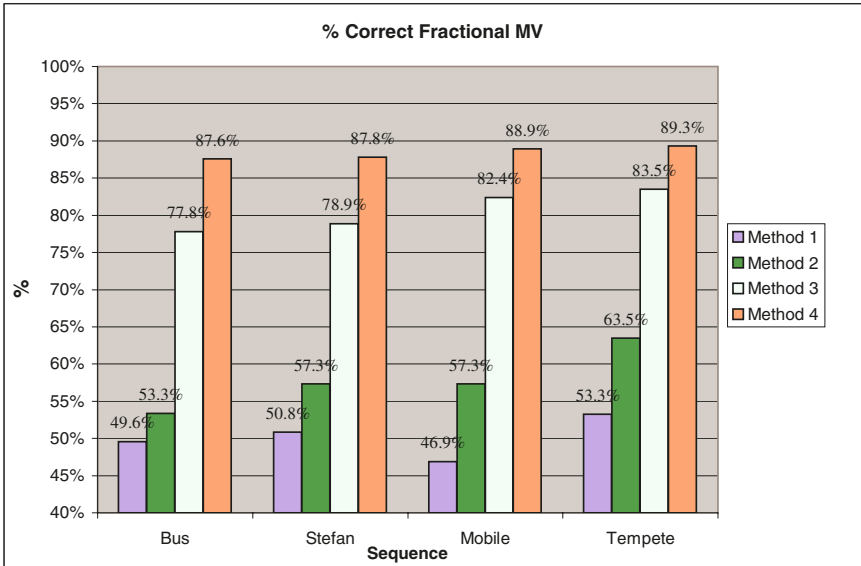


Fig. 8. Percentage of correct fractional MV obtained from different implementations for high-motion sequences

Table 1. Entropy of difference between best fractional MV and optimal MV and average fractional search points for different implementations

Sequence	Measure	Type of Implementation			
		Method 1	Method 2	Method 3	Method 4
Hall	Entropy	0.55	0.48	0.17	0.12
	Ave. Search Pts	1.12	1.64	5.40	5.91
Mother	Entropy	0.64	0.56	0.31	0.23
	Ave. Search Pts	1.23	1.97	5.60	6.19
Container	Entropy	0.45	0.39	0.15	0.09
	Ave. Search Pts	1.14	1.53	5.28	5.77
Coastguard	Entropy	1.01	0.89	0.49	0.34
	Ave. Search Pts	1.63	2.73	6.20	7.45
News	Entropy	0.61	0.54	0.26	0.20
	Ave. Search Pts	1.19	1.72	5.48	6.03
Foreman	Entropy	1.15	1.09	0.76	0.58
	Ave. Search Pts	1.72	2.58	6.32	7.49
Bus	Entropy	1.37	1.32	0.84	0.56
	Ave. Search Pts	1.82	2.52	6.66	8.36
Stefan	Entropy	1.39	1.32	0.82	0.56
	Ave. Search Pts	1.81	2.58	6.63	8.31
Mobile	Entropy	1.38	1.23	0.72	0.52
	Ave. Search Pts	1.84	3.44	6.98	8.61
Tempete	Entropy	1.31	1.14	0.69	0.51
	Ave. Search Pts	1.75	3.41	6.85	8.21

Table 2. Comparison of the fast sub-pixel algorithms with respect to the hierarchical fractional-pel search method (HFPS)

Sequence	No. ref	Algorithms	Measures		
			δ PSNR (dB)	Ave. Pts	δ Bitrate (%)
Hall	2	CBFPS	0.01	5.51	0.15
		Proposed	-0.02	5.37	0.42
Mother	2	CBFPS	-0.05	5.67	-0.44
		Proposed	-0.09	5.52	-0.65
Container	2	CBFPS	-0.01	5.40	-0.07
		Proposed	-0.04	5.24	0.45
Coastguard	3	CBFPS	0.00	6.42	-0.39
		Proposed	-0.02	6.13	-0.14
News	3	CBFPS	-0.02	5.59	0.19
		Proposed	-0.03	5.45	0.42
Foreman	5	CBFPS	-0.03	6.62	-0.08
		Proposed	-0.04	6.21	0.63
Bus	5	CBFPS	-0.01	7.40	0.16
		Proposed	-0.05	6.58	0.91
Stefan	5	CBFPS	-0.01	7.38	0.21
		Proposed	-0.03	6.58	0.70
Mobile	5	CBFPS	-0.02	7.41	0.09
		Proposed	-0.04	6.93	0.48
Tempete	5	CBFPS	-0.01	7.20	0.01
		Proposed	-0.03	6.80	0.36

are presented for the encoding frame-rates of 30 frames per second (fps) and at different quantization accuracy (28, 32, 36, 40).

We also make use of several measures to judge and compare the performance of our algorithm.

1. δ PSNR – average increase in Peak Signal-to-Noise ratio per frame, compared to the HFPS method. A negative value shows a loss of PSNR, compared to the HFPS.
2. Ave. Pts – The average number of fractional search points used for each SPME. The HFPS method uses a fixed 17 fractional search points.
3. δ Bitrate – Percentage of savings in total number of bits needed to encode the sequence, compared to HFPS. A positive value shows an increase in the bits needed for the encoding.

The result presented for each sequence is the average of the results obtained for the different quantization accuracy. The sequences (*Hall*, *Mother*, *Container*, *Coastguard*) is in QCIF format, while the rest of the sequences are in CIF format. The search range used for the motion estimation is ± 32 and ± 16 for the CIF and QCIF sequences respectively. “No.ref” represents the number of reference frames used for the motion estimation. Our algorithm is compared to the CBFPS algorithm and the results are shown in Table 2.

As seen in Table 2, the decrease in PSNR when our proposed algorithm is used is < 0.1 dB, while the encoding bitrate only increase marginally by $< 1\%$, compared to the HFPS method. This corresponds to an overall decrease of PSNR that is limited to 0.1 dB for all bitrates, compared to HFPS. Therefore, our proposed algorithm does not affect the quality of the motion estimation at all. On the other hand, the reduction in fractional search points is much more significant, where there is a reduction of $> 58\%$ in the number of points used, compared to HFPS. This corresponds to an overall savings of 20% in computational time.

Comparing our proposed algorithm to the CBFPS, the quality and encoding bitrate is quite similar. However, our algorithm has the advantage of using less number of fractional search points. Our algorithm uses 0.15-0.8 less search points, compared to CBFPS.

5 Conclusion

In this paper, we propose the use of an adaptive rood pattern and a simplified Small Diamond Search for the sub-pixel motion estimation. The size of the rood pattern is adaptively calculated based on the fractional motion information from adjacent blocks and it is used to obtain a more accurate search center for the refinement search using the simplified Small Diamond Search. The simplified Small Diamond Search is sufficient in obtaining the best fractional MV while retaining a modest number of search points. Our algorithm is able to reduce the number of fractional search points by $> 50\%$, while restricting the PSNR loss to < 0.1 dB, compared to the Hierarchical Fractional-Pel Search (HFPS). The reduction

of these fractional search points correspond to a reduction of $> 20\%$ in computational time. Furthermore, our algorithm has similar quality to the Center-Biased Fractional-Pel Search algorithm, but uses less number of search points.

Acknowledgements

The authors would like to thank the Centre for Design Technology (CDT) Singapore for their support in this research.

References

1. Zhibo Chen, Peng Zhou, Yun He: Fast Integer Pel and Fractional Pel Motion Estimation for JVT. in 6th meeting of Joint Video Team (JVT), JVT-F017 (2002).
2. H.Y. Cheong, A.M. Tourapis: Fast Motion Estimation within the H.264 codec. Proc. of IEEE Int. Conf. on Multimedia and Expo ICME 3(3)(2003) 517–520.
3. Thomas Wiegand, Gary J. Sullivan, Gisle Bjontegaard, Ajay Luthra: Overview of the H.264/AVC Video Coding Standard. IEEE Trans. on Circuits & Systems for Video Technology 13(7) (2003) 560–576.
4. W. Choi, B. Jeon, J. Jeong: Fast motion estimation with modified diamond search for variable motion block sizes. Proc. Int. Conf. on Image Processing ICIP 2(3) (2003) 371–374.
5. Kai-Kuang Ma, Gang Qiu: An improved adaptive rood pattern search for fast block-matching motion estimation in JVT/H.26L. Proc. of the 2003 Int. Symposium on Circuits and Systems ISCAS 2 (2003) 708–711.
6. JVT Reference Software version 10.1: <http://iphome.hhi.de/suehring/tml/download/>

Temporal Error Concealment Based on Optical Flow in the H.264/AVC Standard

Donghyung Kim, Jongho Kim, and Jechang Jeong

Department of Electrical and Computer Engineering, Hanyang University,
17 Haengdang, Seongdong, Seoul 133-791, Korea
{kimdh, angel, jjeong}@ece.hanyang.ac.kr

Abstract. The H.264/AVC standard uses new coding tools to improve coding efficiency. Among the tools, motion estimation using smaller block sizes leads to higher correlation between the motion vectors of neighboring blocks. This characteristic of H.264/AVC is useful for motion vector recovery to conceal a lost macroblock. In this paper, we propose a motion vector recovery method based on optical flow in H.264/AVC video coding. We first determine the optical flow region to alleviate the complexity, and choose an initial value of flow velocity using neighboring motion vectors of a lost macroblock. The proposed method recovers the motion vectors of 4x4 blocks included in a lost macroblock using the weighted average of obtained flow velocities. Simulation results show that our proposed method gives higher objective and subjective visual qualities than conventional approaches.

1 Introduction

When video streams are transmitted through a noisy channel, channel noise or congestion often leads to packet loss. This can drastically degrade the visual quality of the decoded sequence. As a way to solve this problem, error concealment is very useful, since the decoded frame which has lost blocks still includes spatial and temporal redundancy. In temporal error concealment, correlation between the current decoded frame and previous decoded frames is exploited. A damaged macroblock of the current decoded frame is replaced by a macroblock in the reference frame using the estimated motion vector of the lost macroblock.

Among conventional approaches for temporal error concealment, the simplest way is a temporal replacement (TR) method which conceals a lost macroblock with a macroblock located at the same position in a previous frame. A TR method produces reasonably good visual qualities in stationary areas, but significant degradations in dynamic areas.

A boundary matching algorithm (BMA) is one of the most popular methods for motion vector recovery. It exploits the fact that adjacent pixels in a video frame exhibit high spatial correlation. The reference software of H.264/AVC (we called it just H.264) also uses a temporal error concealment method based on BMA.

Several approaches to temporal error concealment have been proposed to enhance performance. Chen et al. proposed the so-called refined boundary matching algorithm

(RBMA), which is based on the boundary matching algorithm [1]. To better satisfy the criterion of minimizing the boundary differences between a lost macroblock and a replaced macroblock in the reference, they use different motion vectors for different regions of a lost macroblock. Zheng et al. proposed a temporal error concealment method for H.264 using the Lagrange interpolation formula to constitute a polynomial that describes the motion tendency of motion vectors [2]. It exploits the fact that there is higher correlation between the motion vectors of adjacent blocks in H.264. Suh et al. proposed the motion vector recovery method using optical flow in MPEG2 [3]. To alleviate the complexity, they determine the optical flow region, and recover the motion vector of a 16x16 block by using the average value of flow velocity vectors of neighboring macroblocks. Xu et al. proposed a set of error concealment schemes to improve the error resilience ability for video consumer applications [4]. Their methods include refined motion compensated temporal concealment with weighted boundary matching criteria, an algorithm of refined directional weighted spatial interpolation, and an adaptive spatial/temporal estimation method with low complexity to combine the above algorithms.

In this paper, we propose a motion vector recovery method for temporal error concealment based on optical flow in H.264. Since optical flow fields are very similar to true motion and can be used to recover two-dimensional motion information, we first compute flow velocity vectors by Horn and Schunck's method [5], and then recover the motion vectors of 4x4 blocks in a lost macroblock using the weighted average of flow velocities.

2 Horn and Schunck's Method

Optical flow is referred to as the two dimensional distribution of apparent velocities of intensity pattern movements in an image plane. In other words, an optical flow field consists of a dense velocity field with one velocity vector for each pixel in the image plane. If we know the time interval between two consecutive images, which is usually the case, then velocity vectors and displacement vectors can be converted from one to another. In this sense, optical flow is one of the techniques used for error concealment.

2.1 Constraints for Determining Optical Flow

To find optical flow, Horn and Schunck use two constraints. They are the brightness invariant constraint and the smoothness constraint.

Brightness Invariance Constraint. Let the image brightness at the point (x,y) in the image plane at time t be denoted by $E(x(t),y(t),t)$. If the image brightness is invariant with respect to the time interval from t to $t+\Delta t$, we then have

$$E(x(t), y(t), t) = E(x(t + \Delta t), y(t + \Delta t), t + \Delta t) \quad (1)$$

Equation (1) is the brightness invariance equation; strictly speaking, it is the brightness time-invariance equation. The expansion of the right-hand side of Eq. (1) in the Taylor series at time t leads to

$$\left(\frac{\partial E}{\partial x}u + \frac{\partial E}{\partial y}v + \frac{\partial E}{\partial t}\right)\Delta t + \varepsilon = 0, \quad u = \frac{dx}{dt}, v = \frac{dy}{dt} \tag{2}$$

where ε contains second and higher order terms, and u and v are the horizontal and vertical components of an optical flow velocity, respectively.

After dividing both sides of the equation by Δt and evaluating the limit as $\Delta t \rightarrow 0$, we have a single linear equation in the two unknown parameters: u and v .

$$E_x u + E_y v + E_t = 0 \tag{3}$$

where E_x , E_y and E_t are the partial derivatives of image brightness with respect to x , y and t , respectively.

Figure 1 depicts the brightness invariance constraint.

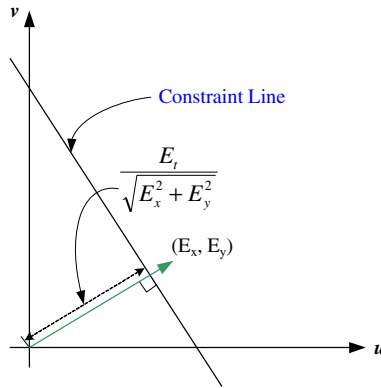


Fig. 1. Brightness invariance constraint

Smoothness Constraint. Equation (3) reveals that we have two unknowns: u and v , but only one equation to relate them. It indicates that there is no way to compute optical flow without an additional constraint.

The smoothness constraint means flow velocity vectors vary from one to another smoothly, particularly for points belonging to the same object. Mathematically, the smoothness constraint is imposed in optical flow determination by minimizing the square of the Laplacians of the x and y components of flow.

$$\begin{aligned} (u, v) &= \arg \min_{u, v} (\nabla^2 u + \nabla^2 v) \\ &= \arg \min_{u, v} \left(\frac{\partial^2 u}{\partial x^2} + \frac{\partial^2 u}{\partial y^2} + \frac{\partial^2 v}{\partial x^2} + \frac{\partial^2 v}{\partial y^2} \right) \end{aligned} \tag{4}$$

2.2 Approximation

For determining optical flow using two constraints, we must compute the partial derivatives (E_x , E_y , E_t in Eq. (3)) and also the Laplacian of the flow velocities ($\nabla^2 u$, $\nabla^2 v$ in Eq. (4)). Horn and Schunck estimate E_x , E_y and E_t at a point in the center of a cube formed by eight measurements, as shown in Fig. 2.

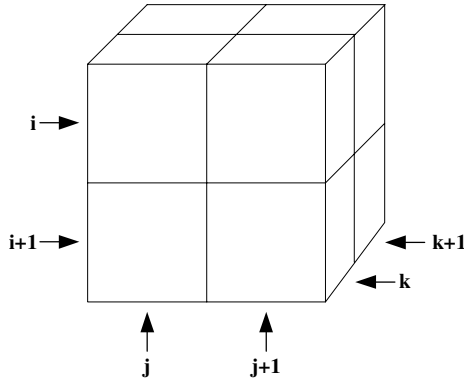


Fig. 2. The cube to estimate the three partial derivatives

Each of the estimates is the average of four first differences taken over adjacent measurements in the cube.

$$\begin{aligned}
 E_x &= \frac{1}{4} (E_{i,j+1,k} - E_{i,j,k} + E_{i+1,j+1,k} - E_{i+1,j,k} \\
 &\quad + E_{i,j+1,k+1} - E_{i,j,k+1} + E_{i+1,j+1,k+1} - E_{i+1,j,k+1}) \\
 E_y &= \frac{1}{4} (E_{i+1,j,k} - E_{i,j,k} + E_{i+1,j+1,k} - E_{i,j+1,k} \\
 &\quad + E_{i+1,j,k+1} - E_{i,j,k+1} + E_{i+1,j+1,k+1} - E_{i,j+1,k+1}) \\
 E_t &= \frac{1}{4} (E_{i,j,k+1} - E_{i,j,k} + E_{i+1,j,k+1} - E_{i+1,j,k} \\
 &\quad + E_{i,j+1,k+1} - E_{i,j+1,k} + E_{i+1,j+1,k+1} - E_{i+1,j+1,k})
 \end{aligned}
 \tag{5}$$

The average can remove the noise effect, thus making the obtained partial derivatives less sensitive to various noises.

The Laplacians of u and v are approximated by

$$\begin{aligned}
 \nabla^2 u &= \bar{u}_{x,y,k} - u_{x,y,k} \\
 \nabla^2 v &= \bar{v}_{x,y,k} - v_{x,y,k}
 \end{aligned}
 \tag{6}$$

where \bar{u} and \bar{v} indicate the local average with respect to the x and y components of flow vectors, respectively, and are estimated as follows:

$$\begin{aligned}
 \bar{u}_{x,y,k} &= \frac{1}{6} (u_{x-1,y,k} + u_{x,j+1,k} + u_{x+1,y,k} + u_{x,y-1,k}) \\
 &\quad + \frac{1}{12} (u_{x-1,y-1,k} + u_{x-1,y+1,k} + u_{x+1,y+1,k} + u_{x+1,y-1,k}) \\
 \bar{v}_{x,y,k} &= \frac{1}{6} (v_{x-1,y,k} + v_{x,j+1,k} + v_{x+1,y,k} + v_{x,y-1,k}) \\
 &\quad + \frac{1}{12} (v_{x-1,y-1,k} + v_{x-1,y+1,k} + v_{x+1,y+1,k} + v_{x+1,y-1,k})
 \end{aligned}
 \tag{7}$$

2.3 Determination of Optical Flow

To determine optical flow, the Horn and Schunck method minimizes a weighted sum of the error in two constraints.

$$\iint ((E_x u + E_y v + E_t)^2 + \alpha^2 (\nabla^2 u + \nabla^2 v)) dx dy \tag{8}$$

In Eq. (8), α^2 plays a significant role only for areas where the brightness gradient is small, preventing haphazard adjustments to the estimated flow velocity occasioned by noise in the estimated derivatives.

Using the calculus of variations and the approximation of the Laplacians shown in Eq. (6), the minimization of Eq. (8) requires solving the two equations in Eq. (9) simultaneously.

$$\begin{aligned} (\alpha^2 + E_x^2 + E_y^2)(u - \bar{u}) &= -E_x(E_x \bar{u} + E_y \bar{v} + E_t) \\ (\alpha^2 + E_x^2 + E_y^2)(v - \bar{v}) &= -E_y(E_x \bar{u} + E_y \bar{v} + E_t) \end{aligned} \tag{9}$$

Finally, as described in Eq. (10), optical flow can be computed iteratively from the estimated derivatives and the average of the previous velocity estimates using the Gauss-Seidel method [6]. In this formula, n is the iteration number.

$$\begin{aligned} u^{n+1} &= \bar{u}^n - E_x(E_x \bar{u}^n + E_y \bar{v}^n + E_t) / (\alpha^2 + E_x^2 + E_y^2) \\ v^{n+1} &= \bar{v}^n - E_y(E_x \bar{u}^n + E_y \bar{v}^n + E_t) / (\alpha^2 + E_x^2 + E_y^2) \end{aligned} \tag{10}$$

3 Proposed Algorithm

3.1 Determining the Optical Flow Region (OFR) and Choosing the Initial Value of the Flow Velocity Vector

Before evaluating the flow velocity vector, we first determine optical flow regions (OFR) which are located at the top, bottom, left and right positions of a lost macroblock. Figure 3 depicts four OFRs used in our method.

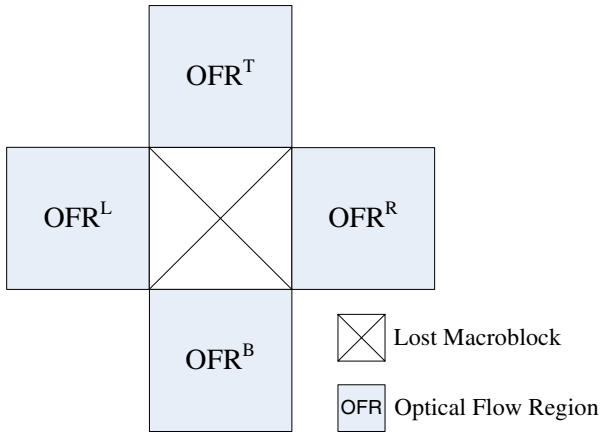


Fig. 3. Optical flow region used in our proposed method

Choosing a more accurate initial value is very important to a fast convergence. Since the H.264 standard estimates the motion vector for finer block size, H.264 has higher correlation between motion vectors of adjacent blocks than previous standards. This characteristic of H.264 enables us to choose a good initial value of a flow velocity vector using motion vectors adjacent to a lost macroblock.

In our algorithm, four OFRs have different initial values. We use the average of motion vectors of four 4x4 blocks adjacent to a lost macroblock as an initial value for each OFR. Figure 4 indicates the motion vectors for choosing the initial value of each velocity vector. Consequently, we can obtain initial values for all OFRs using Eq. (11).

$$(u, v)_{ini}^p = (\sum_{i=0}^3 MV_i^p) / 4 \quad p = T, B, L, R \tag{11}$$

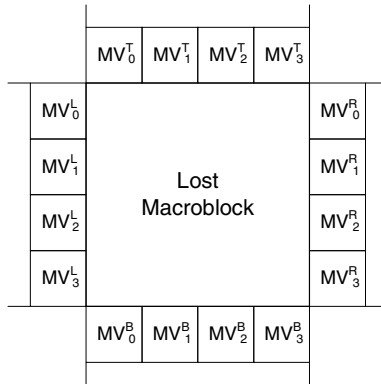


Fig. 4. Motion vectors used for choosing the initial value of the velocity vector

3.2 Motion Vector Recovery

To obtain optical flow of each OFR, we use the iterative scheme described in Eq. (10). After calculation of flow vectors from four OFRs, we take the flow vectors at one-pixel wide outer boundaries and use them to obtain recovery motion vectors of 4x4 blocks. Figure 5 depicts positions of velocity vectors used for motion vector recovery and recovery ordering. In this figure, $(u, v)_i^{T,L,B,R}$ indicates the average velocity vector of four boundary pixels.

For blocks 0 to 3, the process of motion vector recovery is described in Fig. 6. As shown in Fig. 6, the top and left flow velocities are used for motion vector recovery in blocks 0 to 3. In the case of block 0, all weights are equal to 1. In the case of blocks 1 and 2, the different weights are assigned according to distance. The motion vector of block 3 is recovered by using the median vector of blocks 0 to 2.

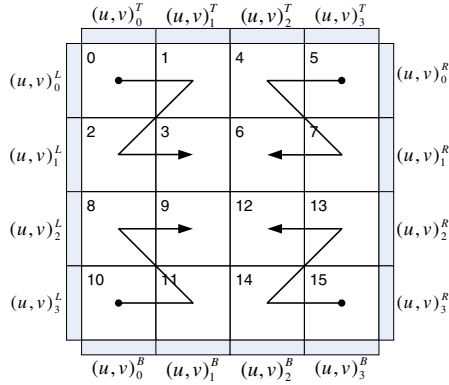


Fig. 5. Velocity vectors used for motion vector recovery and recovery ordering

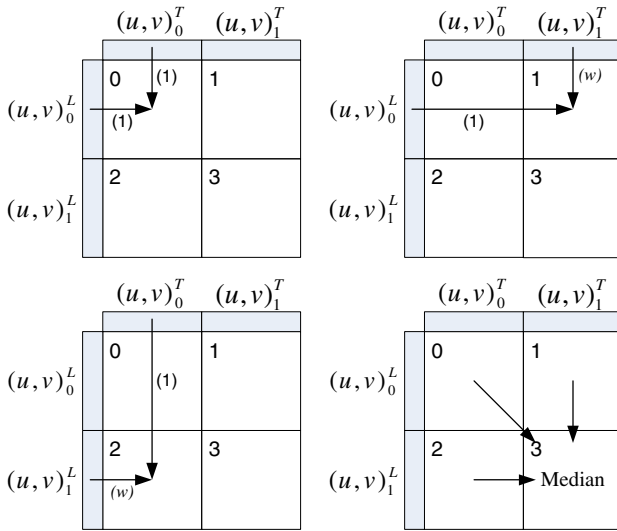
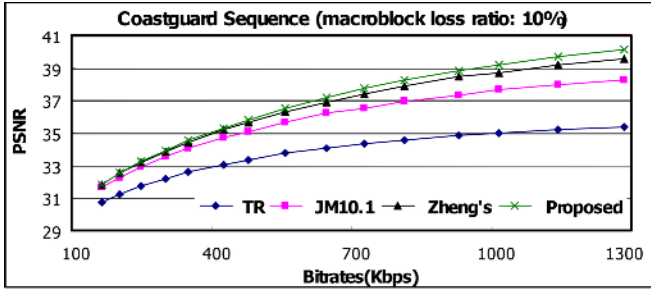


Fig. 6. Motion vector recovery of blocks 0 to 3 ((•) weight)

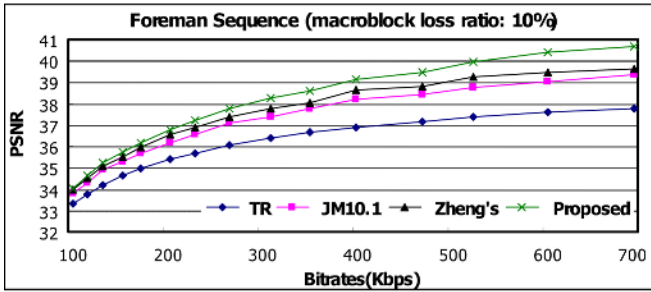
This process can be formulated as Eq. (12).

$$\begin{aligned}
 MV_0 &= ((u, v)_0^T + (u, v)_0^L) / 2 \\
 MV_1 &= (w \cdot (u, v)_1^T + (u, v)_0^L) / (1 + w) \\
 MV_2 &= ((u, v)_0^T + w \cdot (u, v)_1^L) / (1 + w) \\
 MV_3 &= \text{Median}(MV_0, MV_1, MV_2)
 \end{aligned}
 \tag{12}$$

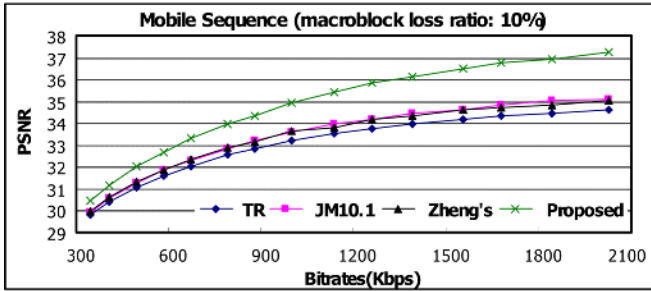
The same process is used for the other blocks (blocks 4 to 15).



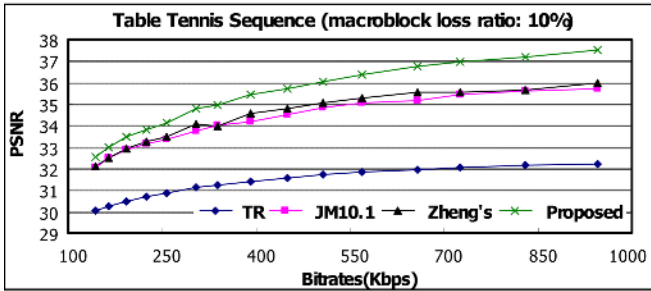
(a)



(b)



(c)



(d)

Fig. 7. Comparison of objective video qualities when there is block error ratio of 10%

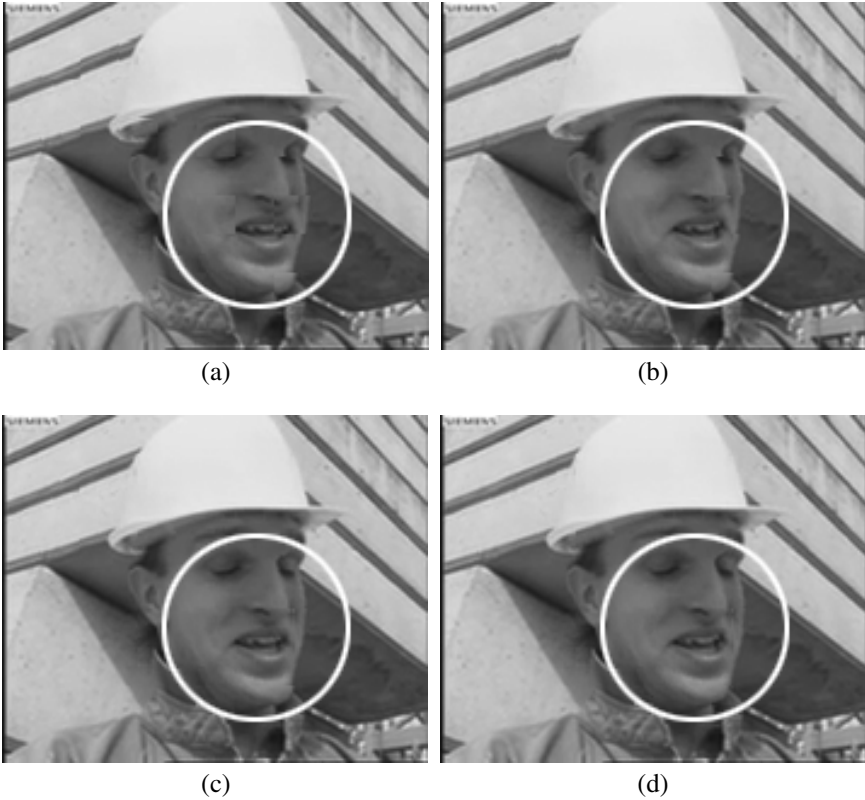


Fig. 8. Comparison of subjective video qualities of the foreman sequence (a) temporal replacement (b) H.264 reference implementation [7] (c) Zheng's method [2] (d) the proposed method

4 Simulation Results

To evaluate the proposed algorithm, we used a public reference encoder, JVT Model (JM) v.10.1 [7]. Four standard video sequences in QCIF (176×144) format were analyzed. These included Coastguard, Foreman, Mobile, and Table Tennis. The first 100 frames of each sequence were used. We compared the simulation results of the proposed algorithm with those of the temporal replacement, the H.264 reference implementation, and Zheng's method [2]. In our simulation, n , iteration times, was limited to be less than 32 in Eq. (10) and w in Eq. (12) was set to 2.

Figures 7 and 8 illustrate objective and subjective qualities of the proposed method compared with conventional approaches. As shown in these results, the proposed method gives better visual qualities than conventional approaches.

5 Conclusions

For temporal error concealment in H.264 we propose the motion vector recovery technique based on optical flow. We first determine four OFRs, estimate the initial value of each OFR from the motion vectors of neighboring blocks adjacent to a lost macroblock, and calculate flow vectors from the initial values. After that, the proposed method recovers the motion vectors of 4x4 blocks in a lost macroblock by weighted averaging of obtained flow vectors. Simulation results show that the proposed method outperforms conventional methods in terms of subjective and objective video qualities.

Acknowledgement

This work was supported by grant No. R01-2003-000-11627-0 from the Basic Research Program of the Korea Science & Engineering Foundation.

References

1. Chen, T., Zhang, X., Shi, Y. Q.: Error Concealment Using Refined Boundary Matching Algorithm. *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 13. (2003) 560-576
2. Zheng, J., Chau, L.P.: A Motion Vector Recovery Algorithm for Digital Video Using Lagrange Interpolation. *IEEE Trans. Broadcasting*, Vol. 49. (2003) 383-389
3. Suh, J.W., Ho, Y.S.: Error Concealment Technique for Digital TV. *IEEE Trans. Broadcasting*, Vol. 48. (2002) 299-305
4. Xu, Y., Zhou, Y.: H.264 Video Communication Based Refined Error Concealment Schemes. *IEEE Trans. Consumer Electronics*, Vol. 50. (2004) 1135-1141
5. Horn, B.K.P., Schunck, B.G.: Determining optical flow. *Artificial Intelligence*, Vol. 17. (1981) 185-203
6. Tekalp, A., *Digital Video Processing*, Prentice Hall (1995)
7. JM 10.1: <http://bs.hhi.de/~suehring/tml/download/jm101.zip> (2005)

Foreground-to-Ghost Discrimination in Single-Difference Pre-processing

Francesco Archetti^{1,2}, Cristina E. Manfredotti²,
Vincenzina Messina², and Domenico G. Sorrenti²

¹ Consorzio Milano Ricerche, via Cozzi 53, 20125 Milan, Italy
`archetti@milanoricerche.it`

² Università Milano-Bicocca, via Bicocca degli Arcimboldi 8, 20126 Milan, Italy
`{archetti, manfredotti, messina, sorrenti}@disco.unimib.it`

Abstract. It is well known that motion detection using single frame differencing, while computationally much simpler than other techniques, is more liable to generate large areas of false foregrounds known as *ghosts*. In order to overcome this problem the authors propose a method based on signed differencing and connectivity analysis. The proposal is suitable to applications which cannot afford the un-avoidable errors of background modeling or the limitations of 3-frames preprocessing.

1 Introduction

Image preprocessing is the first step in many image processing applications such as object localization and tracking [1].

One important preprocessing task is motion detection which is a particularly critical problem since it represents the basic step of information extraction for many complex applications like traffic monitoring [2], video-surveillance [3] and others.

Our aim is to propose a general purpose method for detecting unknown moving objects in a completely unknown environment by satisfying both real-time and low-cost requirements. The only assumption we make is that the unknown background is not in motion.

In the absence of any a priori knowledge about the moving object and about environment the most widely used approaches are background subtraction (e.g. [4], [5], [6], [7],[8], [9]) and frame differencing (e.g. [10], [11]).

Background subtraction bases the detection of moving objects on the difference between the current frame and a reference frame, often called *background* image. This implies that the background image has to be reliable, i.e. it has to be an image of the scene without moving objects. This turns into the need to compute and update a background model, which could account for changes in light conditions or small movements of the scene. A lot of methods have been proposed in the literature on the subject, but no one is reliable in all condition.

If a perfect background modeling would be available, then no other approach could compete with background subtraction.

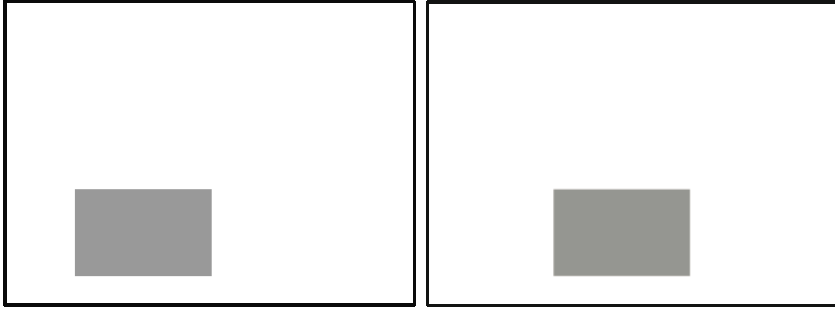


Fig. 1. Left: Image at time $(t - 1)$, I_{t-1} . Right: Image at time t , I_t .

For instance the Median Filter approach, as proposed in [12], [13], [2], [14], computes each pixel of the background image as the average of the corresponding pixels in the n previous images basing on the assumption that a pixel is part of the background image for at least $\frac{n}{2}$ frames, but this assumption is valid only “on average” and not always. Mixture of Gaussians [8], is computationally intensive and its parameters require careful tuning. Moreover, it is very sensitive to sudden changes after a period of stationary conditions [4].

Frame differencing instead uses differences between consecutive images to detect the motion in the actual frame. These approaches present the advantage of using a very up-to-date image of the scene as background. Unfortunately, this image includes the moving objects as well and therefore frame differencing is liable to generate large areas of false foreground, known as *ghosts*. The advantage of frame differencing is that it requires much lower computational effort and avoids errors typically due to the use of a particular background model.

Two kinds of Frame Differencing methods exist in literature: Single Difference and Double Difference. Single Difference uses the video frame at time $(t - 1)$ as the background image for the frame at time t .

Let $I_t(x, y)$ be the pixel intensity at the spatial location (x, y) and \bar{T} a given threshold parameter, then pixels detecting motion are given by:

$$SDiff_t(x, y) = \begin{cases} |I_t(x, y) - I_{(t-1)}(x, y)| & \text{if } |I_t(x, y) - I_{(t-1)}(x, y)| > \bar{T} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

and the corresponding Motion Image, $IM_t(x, y)$ can be defined as follows:

$$IM_t(x, y) = \begin{cases} 1 & \text{if } SDiff_t > 0 \\ 0 & \text{otherwise.} \end{cases} \quad (2)$$

Notice that $IM_t(x, y)$ includes both the moving objects, i.e. the foreground, and their ghosts (see Fig. 2 left).

The ghost problem is known to be solvable, with some disadvantages, by using the Double Difference method [10], which uses $I_{(t-2)}$ as the background image for the frame at time $(t - 1)$, and the latter as the background image for the frame at time t . In this case the motion is detected in image $DDiff_t$ given by:

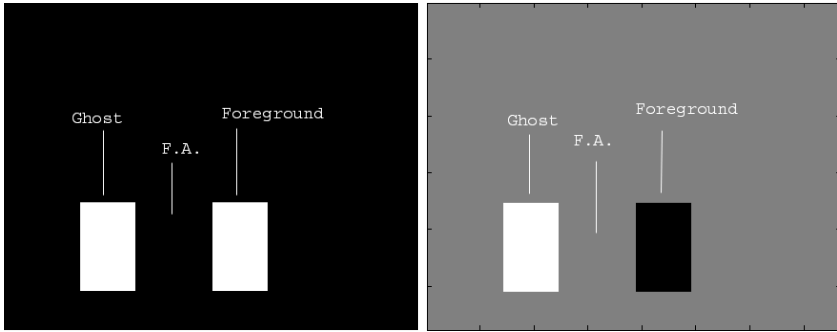


Fig. 2. Left: SDiff: both foreground pixels and ghost pixels are set to 1 in the Motion Image shown. Right: SSDiff: ghost and foreground pixels have different intensity.

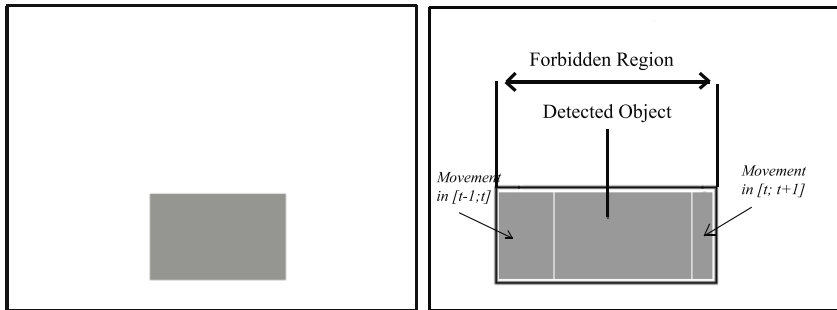


Fig. 3. Left: Image at time $t + 1$, $I_{(t+1)}$. Right: DDiff:the ghost problem is solved. The image represents the forbidden region.

$$DDiff(x, y) = IM_{(t-1)}(x, y) \text{ AND } IM_t(x, y).$$

As mentioned above, this method solves the ghost problem of the Single Difference (see Fig. 3 right), but introduces a minor and a major drawback:

1. The minor drawback is the introduction of one frame delay in the detection of moving objects, i.e. at time t we detect what was moving at time $(t - 1)$;
2. As in any frame difference based approach, a moving object A spans part of the image space during time: this area is *forbidden* to other moving objects because the detection of other moving objects in this area would be mixed up with the detection of A . The major drawback is the increasing of the *forbidden area* with the time interval $[(t - 2), t]$, given the frame rate. Note that its weightless is also counter-proportional to the speed of the target. In order to handle a higher speed, one could either increase the frame rate, which could be costly, or take the background difference route, which requires background modeling.

Our proposal tries to circumvent the ghost problem for the Single Difference approach, and could be of interest in many applications which cannot afford the limitations of Double Differencing or the errors introduced by the background modeling methods or their computational cost.

The paper is organized as follows: in section (2) we present our approach, in section (3) we present experimental evidence of the effectiveness of our proposal. In section (4) conclusions and future works are presented.

2 Our Approach

The method we propose is based on the difference between consecutive frames. Usually, such difference is calculated as the absolute value of the difference in the intensity, as in (1). In *SDiff* motion is detected in two areas, one due to the image position the object had at time $(t - 1)$ (ghost), and the other due to the object position in the current frame (foreground). The two instances have similar image intensity in *SDiff*, and the possibility of distinguish between the two is lost.

In Fig. (2), another relevant area, called *Foreground Aperture*, is emphasized. It is formed by the overlapping of the target positions in $I_{(t-1)}$ and I_t and its weightless depends by the motion of the target in the image with respect to the frame rate. The foreground aperture (*F.A.*) area is characterized by pixels where the intensity is close to zero, due to the effect of the subtraction of pixels more or less at the same intensity level: in this area no motion can be detected.

We propose to use the Signed Single Difference and to separate pixels of positive intensity from pixels of negative intensity: this separation will be used to discriminate the foreground from the ghost, as shown in Fig. (2) right. In order to discriminate between the two we propose a heuristic whose generality, we claim, is quite high.

Given two consecutive frames, $I_{(t-1)}$ and I_t (Fig. 1, left and right) we consider the Signed Single Difference, $SSDiff_t(x, y)$:

$$SSDiff_t(x, y) = \begin{cases} (I_t(x, y) - I_{(t-1)}(x, y)) & \text{if } |I_t(x, y) - I_{(t-1)}(x, y)| > \bar{T} \\ 0 & \text{otherwise} \end{cases} \quad (3)$$

2.1 The Heuristic

We are now left with some ghost and some foreground areas, which we want to discriminate. In Fig. 4 left a simple example with one foreground and one ghost area is shown.

The basic idea is that around both the foreground and ghost area, as detected in the $SSDiff_t(x, y)$ image, we have background in $I_t(x, y)$ (see Fig. 4 right). The heuristic can base on the different behavior/mode of the image intensity pattern inside and outside the corresponding areas in $I_t(x, y)$. In other words, if some local descriptor changes significantly, when computed inside the supposed-foreground and in some neighborhood of it, in the current image $I_t(x, y)$, we can

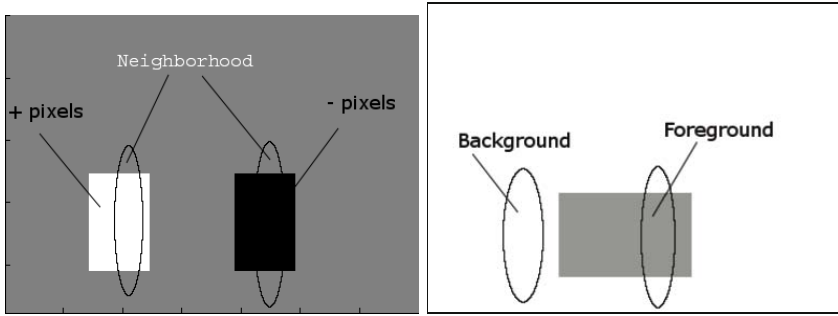


Fig. 4. Left: The neighborhood for the heuristic are defined in the $SSDiff$ image. Right: The descriptor for the neighborhood are evaluated in the current image, I_t .

deduce that the supposed-foreground is indeed foreground. Conversely, doing the same for the ghost area would turn into no significant change for the descriptor.

We propose to segment $SSDiff_t$ in regions, by applying an usual connectivity analysis; this allows the definition of the areas where to apply the heuristic. We can then work on a neighborhood of the blobs (B_i) in the current image $I_t(x, y)$. The algorithm we propose is therefore built of the following steps:

1. $SSDiff_t$ is divided in two images containing the pixels of positive and negative value respectively.
2. The two images are segmented and several blobs (B_i) are identified; note that, the sign of the single difference becomes useful here as it allows the discrimination into different blobs.
3. Compute the descriptor value for each B_i from I_t .
4. Compute the descriptor value for the neighborhood of each B_i , denoted as $n(B_i)$, from I_t .
5. classify B_i as foreground if $descriptor(B_i)$ is significantly different from $descriptor(n(B_i))$; classify B_i as ghost otherwise.

Concerning the descriptors, which can be used to discriminate the nature (foreground or background/ghost) of the different B_i , there are many possible choices, some of which are mentioned below:

1. the average light intensity,
2. the average standard deviation of the light intensity,
3. some texture descriptor, as a parameter of some statistical modeling in spatial domain

We can choose one of this descriptor or a weighted sum of them. We are not interested here in the specific pros and cons of each of them, as we are presenting the approach.

Our claim is that the value of a suitable descriptor takes significantly different values when computed in B_i , with respect to the value taken in $n(B_i)$, if B_i is part of the foreground; we recall that all values are computed in the current image, I_t

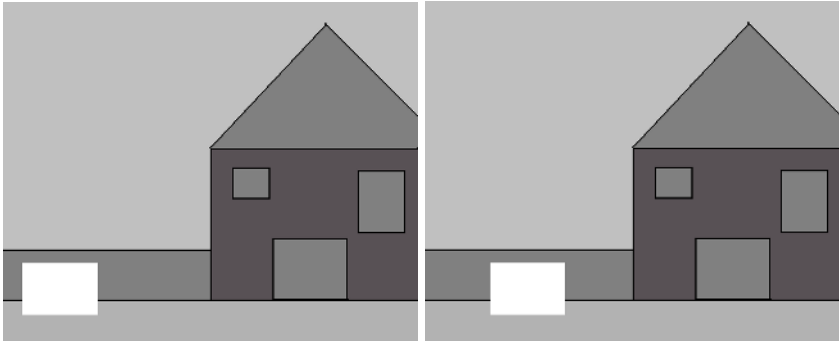


Fig. 5. Left: The image at time $t-1$. Right: The image at time t in a textured-toy-world.

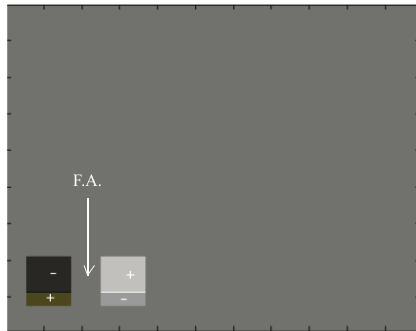


Fig. 6. The SSDiff, the textured background make the sign of blobs of the same area different

but the B_i (and the $n(B_i)$ as well) are detected in $SSDiff_t$. Conversely, if the descriptor value does not change significantly, when moving from B_i to $n(B_i)$ it means we are dealing with a background area.

We now move from the toy world (flat image intensity) used to introduce the approach, to more realistic textured images. Texture of the target is not a problem because what would happen is that the target would split into blobs of different sign. This is not a problem because each of the blob would then be classified as foreground.

If the background is textured, both the target and ghost split into blobs of different sign. We need to consider the scale of the texture patches.

If the texture patches are much smaller than the foreground ones, there is no problem, provided a suitable descriptor is chosen. If the texture patches are comparable in size with the foreground ones, we can only report, for some of the blobs of the target and of the ghost, the compatibility with both being foreground and being background. This is not a failure, but a situation in which we have lack of information. The amount of blobs for which the ambiguity will not be resolved depends on the shape of the target and the texture patches.

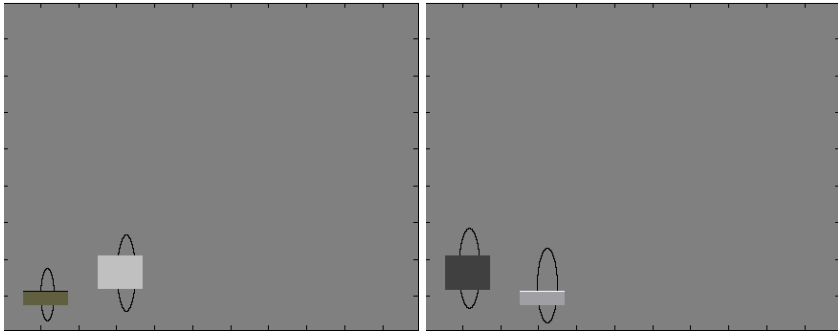


Fig. 7. Left: The neighborhood for the heuristic are defined in the two images, positive pixels. Right: negative pixels.

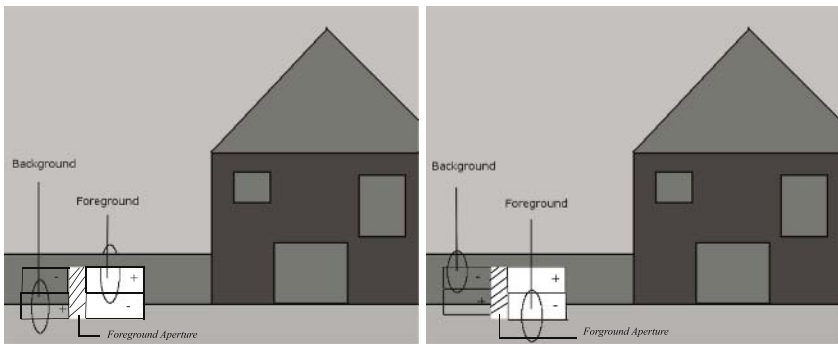


Fig. 8. Left: The neighborhood for the heuristic are used in the current image I_t , positive pixels. Right: negative pixels, even in this case it is possible to correctly detect the ghost and the foreground area. With “+” and “-” the positive and negative areas are shown.

Given two consecutive frames (Fig. 5), we consider the Signed Single Difference (Fig. 6) and we separate it into images depending on the sign of the value of its pixels (Fig. 7). We consider the neighborhood in the frame at time t (Fig. 8): the heuristic discriminates between ghost and foreground even in presence of B_i of different sign.

The complexity of this method is strictly dependent on the number of blobs in the image. Processing a scene as in Fig. 9 a code that needs to be optimized takes 60 ms.

3 Experiments

We present some illustrative examples, taken from our main application domain, which is traffic monitoring. The first one deals with many moving objects. Fig. (9) shows a typical indoor scenario with i.e. three cars in a tunnel.

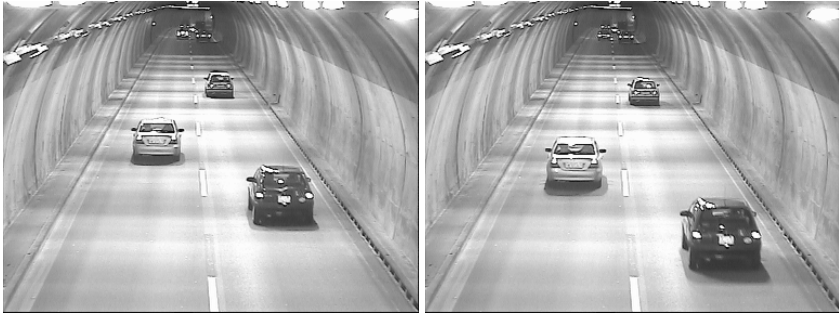


Fig. 9. Left: Image at time t , I_t . Right: Image at time $(t - 1)$, I_{t-1} .

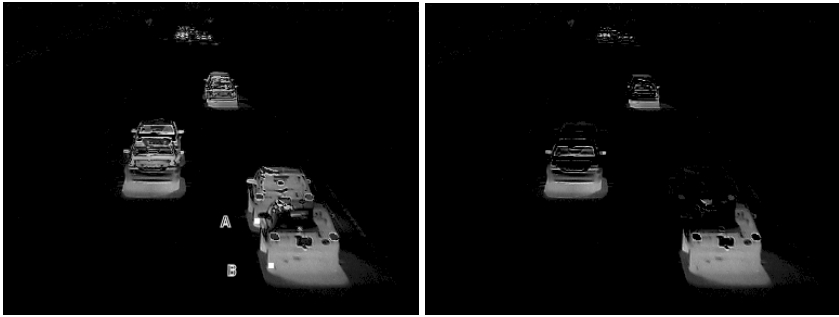


Fig. 10. Left: Absolute Single Difference. Right: Signed Single Difference.

By applying Single Difference we obtain the image shown in Fig. (10) on the left, where one can notice that point A has the same intensity of point B (grey level 160), even though they belong to two different areas (foreground and ghost).

Instead, the result of applying Signed Single Difference is shown in Fig. (10) on the right, where the foreground and the ghost are clearly more distinguishable. This is even more evident in Fig. (11) left and right, where the pixels of positive and negative value are shown respectively.

In order to discriminate between ghost and foreground we apply the proposed heuristic. We first perform the connectivity analysis [15] by the usual 8-con algorithm. Blobs having an area less than a given threshold, 10 pixels in this experiment, have been thrown away. The overall result is shown on the left of Fig. (12).

We come now to the definition of the neighborhood of a blob. Many alternatives are available and we are not interested, as we are presenting the approach, in the most efficient one. In these experiments we considered the neighborhood as the area covered by widening the blob bounding box by 5 pixels. In Fig. (12) on the right we show the blobs altogether with their bounding boxes. In Fig. (13), on the left, we show the neighborhood of the detected blobs used, super-

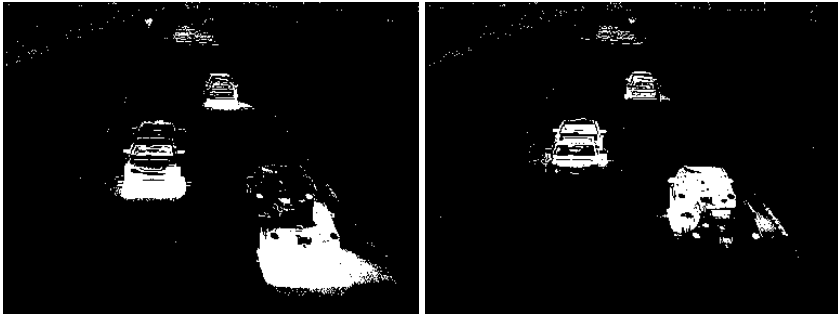


Fig. 11. Left: Image of the positive pixels only. Right: Image of the negative pixels only.

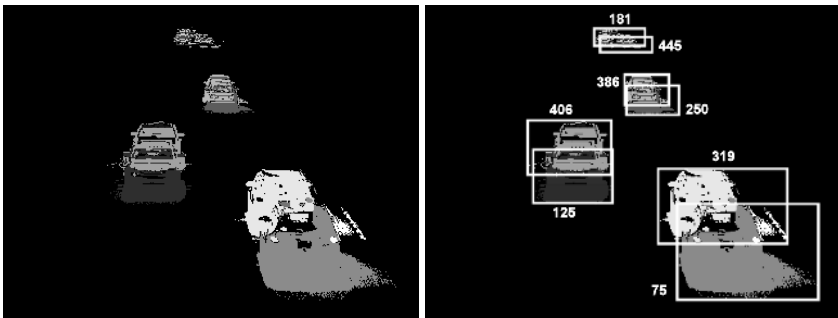


Fig. 12. Left: Each blob is depicted in a different level of grey. Right: Blobs and their bounding boxes; the numbers are the blob labels.

imposed to the current image. On the right, we show the blobs which have been classified as foreground.

In these experiments the descriptor we adopted is the simple average image intensity. We are therefore interested in the descriptor values inside each blob and outside its bounding box (the area 5 pixels around it). In Table (1) we report the differences that this descriptor takes inside and outside the blobs. Column 2 reports the labels given by the connectivity analysis to each blob, while in column 1, one can find the computed differences, in an increasing order. The ghosts are identified by choosing the first four values of differences, i.e. the smallest ones. The remaining four are classified as foreground.

We present hereafter other examples of significant situations, whose appropriate handling is obviously required. We present video footage for a fast target together with a large slow target in indoor, and for many overlapping targets in outdoor. For each situation we show I_t , $I_{(t-1)}$ and the blobs classified as foreground. As for the experiment described above, we used the (simple) average image intensity as descriptor and the area 5 pixels around the bounding box of each blob as neighborhood. Notice in the first example, Fig. (14), that the



Fig. 13. Left: Image at time t showing the bounding boxes and the considered neighborhood. Right: Blobs classified as foreground.

Table 1. Differences of the average luminance between inside the blobs and outside their bounding box

differences	labels
0.0669	125
18.0231	75
18.5301	250
21.6420	445
81.6571	181
83.6615	386
98.2408	319
102.1331	406



Fig. 14. Left: Image at time t , I_t . Center: Background, $I_{(t-1)}$. Right: Blobs classified as foreground.

preprocessing detects both targets. In the second example, Fig. (15), all targets are detected, apart the fifth, which is very far and has nearly no image motion. The interesting point is that the two nearest targets have ghost and foreground mixed up, but still the heuristic, even with this simple selection of neighborhood and descriptor, is capable to distinguish them.



Fig. 15. Left: Image at time t , I_t . Center: Background, $I_{(t-1)}$. Right: Blobs classified as foreground.

4 Conclusions

We presented a proposal for overcoming the ghost problem in Single Difference preprocessing. The proposal is based on signed difference and connectivity analysis, and is suitable to applications which cannot afford the limitations of 3-frames preprocessing, like the double difference, or the errors of reliable background modeling or its computational cost. The results obtained with experimentation is promising. Further investigations are ongoing about the applicability of similar ideas to the detection of the foreground aperture area.

References

1. Amamoto, N., Fujii, A.: Detecting obstructions and tracking moving objects by image processing techniques. *Electronics and Comm. Japan, Part 3* **82** (1999) 28–37
2. Gloyer, B., H.K.Aghajan, K.Y.S., T.Kailath: Video-based freeway monitoring system using recursive vehicle tracking. In: *Proceedings of SPIE*. (1995) 173 – 180
3. McKenna, S., Jabri, S., Duric, Z., Wechsler, H.: Tracking interacting people. In: *4th Int. Conf. on Automatic Face and Gesture Recognition, Grenoble, France* (2000) 384–353
4. Cheung, S.C., Kamath, C.: Robust techniques for background subtraction in urban traffic video. In: *Video Communications and Image Processing, SPIE Electronic Imaging, San Jose* (2004)
5. Cheung, S.C., Kamath, C.: Robust background subtraction with foreground validation for urban traffic video. *EURASIP Journal on Applied Signal Processing* **14** (2005) 1–11
6. Kim, K., Khalidabhongse, T.H., Harwood, D., Davis, L.S.: Real-time foreground-background segmentation using codebook model. *Real-Time Imaging* **11** (2005) 172–185
7. A.Elgammal, Duraiswami, R., Harwood, D., L.S.Davis: Background and foreground modelling using non-parametric kernel density estimation for visual surveillance. *Proc. of IEEE* (2002)
8. N.Friedman, S.Russell: Image segmentation in video sequences: a probabilistic approach. In: *Proc. Thirteenth Conf. on Uncertainty in Artificial Intelligence (UAI 97)*. (1997)

9. Stauffer, C., Grimson, W.: Learning patterns of activity using real time tracking. *IEEE Trans. Pattern Analysis and Machine Intelligence* **22** (2000) 747–757
10. Yoshinari, K., Michihito, M.: A human motion estimation method using 3-successive video frames. In: *Proc. of Int.Conf. on Virtual Systems and Multimedia (GIFU)*. (1996) 135–140
11. Zhang, C., Chen, S., Shyu, M., S.Peeta: Adaptive background learning for vehicle detection and spatio-temporal tracking. In: *Information, Communications and Signal Processing*. (2003)
12. R.Cutler, L.Davis: View-based detection. In: *Proceedings Fourteenth International Conference on Pattern Recognition, Brisbane, Australia* (1998) 495–500
13. R.Cucchiara, M.Piccardi, A.Prati: Detecting moving objects, ghost, and shadows in video streams. *IEEE transactions on Pattern Analysis and Machine Intelligence* (2003) 1337 – 1342
14. Q.Zhou, J.Aggorwal: Tracking and classifying moving objects from videos. In: *Proceedings of IEEE Workshop on Performance Evaluation of Tracking and Surveillance*. (2001)
15. Gonzales, R.C., Woods, R.E.: *Digital Image Processing*. Addison-Wesley Publishing Company (1993)

Moving Object Removal Based on Global Feature Registration

Soon-Yong Park¹, Jaekyoung Moon², Chang-Joon Park³, and Inho Lee³

¹ Computer Engineering Department, Kyungpook National University, Daegu
702-701, South Korea

² Sensor Technology Research Center, Kyungpook National University, Daegu
702-701, South Korea

³ Digital Actor Research Team, Digital Content Division, ETRI, Daejeon 305-350,
South Korea

Abstract. A moving object in a video sequence is removed and corresponding background is completed by using a novel global feature registration technique. To find a 2D homography between two adjacent video frames, we track background and foreground features, separately. After estimating the homography, we extract and remove the moving object in every frame. To fill the background of the removed object accurately, we introduce a global feature registration technique. The technique iteratively reduces and distributes the accumulation errors associated to global video registration. Experimental results show that the proposed technique yields seamless background sequences.

1 Introduction

Removing moving objects in a video sequence is one of the digital post-processing techniques. Some common digital post-processing techniques include composition of graphic objects, removal of unnecessary objects, and insertion of real or virtual objects to a video sequence. Until today, most of these tasks are done by user intervention using specially designed graphics software. Thus post-processing is a very time-consuming and difficult job. Recently, several investigations of automatic post-processing are reported which employ computer graphics and computer vision techniques.

This paper presents a computer vision technique of removing moving objects in a video sequence. We track, remove and replace a moving foreground object (hereafter object) with its corresponding background objects (hereafter background). After tracking the object using a feature-based technique, we obtain the background of the object from one of the other frames by accurately estimating the 2D homographic transformation. In Figure 1(a) a moving object is shown in one of the video frames. The object is removed and its empty area is filled with the corresponding background as shown in Figure 1(b).

Recent investigations of object removal fall in two categories, one is filling the background of static objects and the other is of moving objects. Here are



Fig. 1. Object removal example. Left: Original frame. Right: After object removal.

some literature reviews of moving object removal. A. Agarwala et al. [AG] semi-automatically extract and track the boundary features of a moving object. They define keyframes to refine the boundary of the object and generate a new video sequence. Y. Zhang et al. [ZH] separate foreground and background objects in multiple layers to extract a moving foreground object. This layered approach removes a foreground object whether it moves or not. Y. Sugaya and K. Kanatani [SU] assume a feature point is an affine transformation of a 3D feature, then generate a trajectory vector by stacking 2D coordinates of tracked features from all frames. By affine space fitting, they remove outliers, extract and remove a moving object.

Y. Wexler et al. [WE] addresses a problem of completing a moving object which is partly occluded in a video sequence. They use a stationary camera to obtain the video frame. They complete the occluded object areas by filling empty holes with space-time patches which have spatio-temporal consistency. P. Sand and S. Telle [SA] use two video sequences to match features frame by frame. They register two video sequences and apply special effects such as object removal, object insertion, and rotoscopy. K. Bhat [BH] extracts a moving object from a video sequence by making a video mosaic of a scene. A. Yamashita et al. [YA] remove noises in a dynamic scene. They change the direction of a pan/tile camera to identify and remove the noises. Their approach can be applied to only adherent noises close to a video camera.

To obtain a new seamless background video sequence, it is necessary to estimate the 2D homographies of all video frames. If the 2D homographies of all video frames are very accurate, we can obtain the seamless video by filling the removed object areas using the matching background. Therefore, finding accurate transformations of all video frames is very important in object removal. In this paper, object removal is done based on a global feature registration technique. We track object and background features frame by frame. Then 2D homographies of every pair of frames are estimated. The RANSAC technique is used to remove erroneous outliers in object and background features. Empty object areas are finally completed by their background images. Figure 2 shows the overview of the proposed technique.

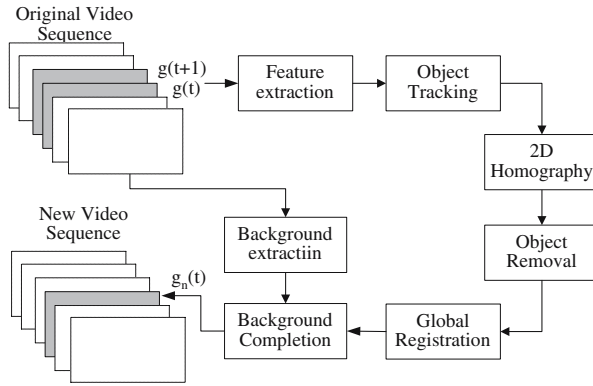


Fig. 2. Overview of the proposed object removal technique

2 Background and Object Tracking

To track a moving object with respect to a moving background, we need to know the motion of the background. Because both object and background are moving, we need to separate the motion of two subjects. In this paper, we track the corner features of object and background separately. Background features are then used to estimate the 2D homographies between video frames, and object features are used to estimate the motion of the object.

2.1 Background Tracking

Suppose we obtain a video sequence of a moving object using a rotating video camera. In the video sequence, we know that both background and foreground objects are moving but in different motion. In the background, we assume there is enough features by which we can track using an object tracking technique. Let \mathbf{f}_B and \mathbf{f}_O be the sets of background and object features. To obtain \mathbf{f}_B and \mathbf{f}_O , we extract and track corner features frame by frame using the KLT (Kanade-Lucas-Tomasi) tracker.

To separate the background and foreground objects, we define two tracking windows, W_O and W_S . W_O is the object window and W_S is the search window and we assume W_O is always inside W_S . In the first frame, the object window is defined manually by a human intervention. We modified the KLT algorithm to find background corners only in W_S . A binary image mask is used to modify the KLT algorithm to find and track corner features only in W_S . After all features are extracted, we regard those in W_S as elements of \mathbf{f}_B and those in W_O as elements of \mathbf{f}_O . However, it needs to determine if a feature in W_O is an object or a background feature. If the feature is determined as a background one, we move it to \mathbf{f}_B . Figure 3 shows two tracking windows, the green (light gray) rectangle is the background window and the red (dark gray) rectangle is the object window. In each window, corner features are shown as dot clouds.



Fig. 3. Object and background tracking windows

2.2 2D Perspective Transformation

Once we find both object and background features between two adjacent frames, i.e. we estimate the 2D perspective transformation (homography) between the two frames. Because the KLT tracker yields some erroneous results, we use the RANSAC algorithm to remove outliers. We assume that the camera is rotating along the vertical axis passing the optical center of the lens (but not exactly) and consider 2D homography as the estimation model of the RANSAC algorithm.

Suppose there are two images $g(t)$ and $g(t - 1)$ which are obtained at time t and $t - 1$, respectively. If the coordinates of two matching points between two images are (x, y) and (x', y') , the 2D perspective transformation matrix M is defined as

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} u \\ v \\ w \end{pmatrix} = M \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} m_{00} & m_{01} & m_{02} \\ m_{10} & m_{11} & m_{12} \\ m_{20} & m_{21} & m_{22} \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \quad (1)$$

By setting $m_{22} = 1$, the above equation is rewritten as

$$x' = \frac{m_{00}x + m_{01}y + m_{02}}{m_{20}x + m_{21}y + 1} \quad (2)$$

$$y' = \frac{m_{10}x + m_{11}y + m_{12}}{m_{20}x + m_{21}y + 1} \quad (3)$$

Because the number of matching features is usually more than the minimum requirement for computing the matrix, we use the linear least squares method to solve the over-determined equation. We estimate the homography so that the registration error of more than 90% of features is within 0.5 pixel.

2.3 Object Tracking

As mentioned in the previous section, we define the object window W_O in the first frame manually by user intervention so that a moving object is inside the

window. Let $O(x, y, t)$ be the coordinates of the upper-left corner of the window $W_O(t)$. The width and the height are assumed constant. If we know $M_{t-1,t}$, the transformation of two video frames obtained at t and $t - 1$, we can write an equation

$$O'(x, y, t) = M_{t-1,t}O(x, y, t - 1) \tag{4}$$

As mentioned in an earlier section, $M_{t-1,t}$ is derived from the background features. Thus $O'(x,y,t)$ is not exact coordinates of $W_O(t)$. However, $O'(x,y,t)$ is very close to $O(x,y,t)$ because the object motion between two adjacent frames is small. Therefore, it needs to determine $W_O(t)$ using the tracking features inside $W_{O'}(t)$. $W_O(t)$ is determined by the following steps.

1. Suppose an object feature $f_O(t - 1)$ moves to $f_O(t)$
2. If a registration error $= |f_O(t) - M_{t-1,t}f_O(t - 1)|$ is less than a predefined threshold value, then $f_O(t)$ is regarded as $f_O(t) \in \mathbf{f}_B$, because $M_{t-1,t}$ is derived from the background features.
3. Else $f_O(t) \in \mathbf{f}_O$, because it is not consistent with the background motion.
4. The center of $W_O(t)$ is decided by the centroid of \mathbf{f}_O .

When the coordinates of the centroid of \mathbf{f}_O at time t is (x_t, y_t) and at time $t - 1$ is (x_{t-1}, y_{t-1}) , motion of the object $v_m(t) = (x_t, y_t) - (x_{t-1}, y_{t-1})$.

In addition to tracking corner features, we also use the difference image of $W_O(t)$ and $W_O(t - 1)$ to obtain another motion vector. To obtain the difference image, two video frames need to be registered to a reference frame to compensate the background motion. By considering $W_O(t)$ as the reference frame, we get an difference image

$$W_d(t) = |W_O(t) - W'_O(t - 1)|, \tag{5}$$

where

$$W'_O(t - 1) = M_{t-1,t}^{-1}W_O(t). \tag{6}$$

$W'_O(t - 1)$ is the transformation of $W_O(t)$ to the frame at $t - 1$. $W_d(t)$ is then binarized by a predefined threshold. Similarly, $W_d(t - 1)$ is obtained between $W_O(t - 1)$ and $W_O(t - 2)$ and a motion vector $v_d(t)$ between two frames is obtained by using the binarized images, $W_d(t)$ and $W_d(t - 1)$. To filter out image noise in the difference image, an accumulation image W_a of consecutive frames is employed [TI]. The relationship between $W_a(t)$ and $W_d(t)$ is as follows:

$$\begin{aligned} W_a(t) &= (1 - w_a)W_a(t - 1) + w_a |W_O(t) - W'_O(t - 1)| \\ W_d(t) &= \begin{cases} 1 & \text{if } (W_a(t) > T_d) \\ 0 & \text{otherwise,} \end{cases} \end{aligned} \tag{7}$$

where w_a and T_d are a weighting factor and a threshold value, respectively. Object motion based on the accumulation image is determined by the centroid difference of $W_d(t)$ and $W_d(t - 1)$. From two motion vectors, we compute the object motion at t as

$$v(t) = w_m v_m(t) + (1 - w_m) v_d(t). \tag{8}$$

In the above equation, w_m is the weighting factor of $v_m(t)$ to determine $v(t)$. It is usually set to 0.5.

3 Background Completion

3.1 Background Matching

If a moving object is tracked successfully in all video frames, we can remove the object regions in the frames. The next step is then to fill the empty object areas with appropriate background images, which is called 'background completion'. Because the 2D homography for every pair of video frames is known already, we can acquire the background image of a removed object from one of the other frames. Suppose the object region $W_O(t)$ needs to be filled by the matching background image. Then the transformation of this region to $g(t+k)$ frame is written as

$$W'_O(t+k) = M_{t,t+k}W_O(t). \tag{9}$$

If $W'_O(t+k)$ does not overlap with $W_O(t+k)$, $W'_O(t+k)$ can be used to fill the empty region of $W_O(t)$. In Figure 4 for example, $W'_O(t+k)$, the transformation of $W_O(t)$ from $g(t)$ to $g(t+k)$ frame, does not overlap with $W_O(t+k)$. However, $W'_O(t-k)$ in $g(t-k)$ frame overlaps with $W_O(t-k)$. Therefore, the empty object region $W_O(t)$ is filled by warping the image of $W'_O(t+k)$. Transformation between $g(t)$ and $g(t+k)$ frames is obtained by multiplying pair-wise transformations from $M_{t,t+1}$, to $M_{t+k-1,t+k}$ such that

$$M_{t,t+k} = M_{t+k-1,t+k} \cdots M_{t+1,t+2}M_{t,t+1}. \tag{10}$$

3.2 Global Registration Refinement

Multiplication of many pair-wise transformations accumulates a significant registration error between source frame and target frame. Without using any registra-

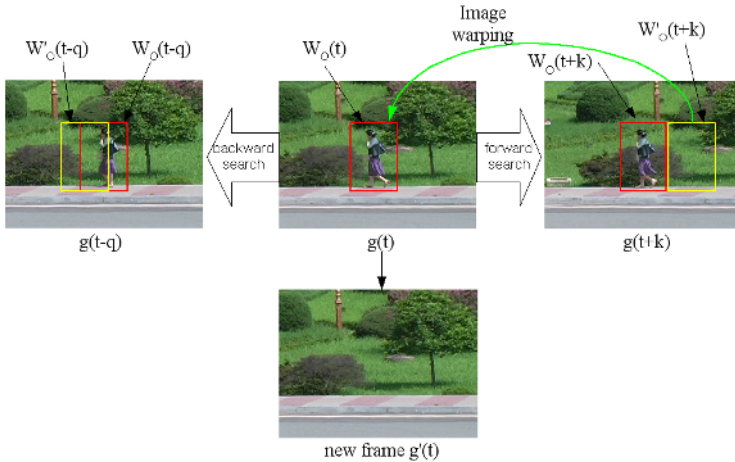


Fig. 4. Matching background search for filling object area

tion refinement technique, background completion could results in visible seams along the boundary of the object window. To solve this problem, we refine all pair-wise transformations between all frames using a global feature registration technique.

The proposed refinement technique iteratively reduces registration error in all frames. The main idea is using the traces of background features to refine the pair-wise registration. Suppose there is a background feature point $f_B(x, y, t)$ at time t . If it is successfully tracked until the $(t + k)th$ frame, i.e. $f_B(x, y, t + k)$, we know the trace of the feature from t to $t + k$. Similarly if we know the traces of all features from $\mathbf{f}_B(t)$ to $\mathbf{f}_B(t + k)$, we can generate a 2D trace table using a 2D linked-list or array. Figure 5 visualizes a 2D form of all feature traces using pseudo colors. The horizontal axis corresponds to the frame number and the vertical axis corresponds to the feature number in each frame. This figure shows all features in the video sequence shown in figure 1. The sequence consists of 150 frames and maximum 500 features are extracted in each frame. However, only about 200 features are successfully tracked in each frame. In the figure, different features are represented with different colors. In a single horizontal line, it is shown that the line color changes several times. A Black-colored lines means the corresponding features fail to track in that time interval.

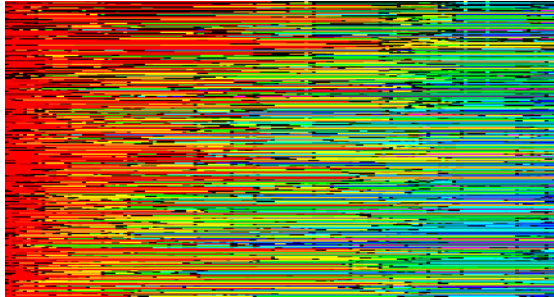


Fig. 5. Tracking feature color table

In Figure 6, a background feature $f_B(x, y, t - 1)$ is tracked across multiple frames from time t to $t + k$. A pair-wise transformation $M_{t-1,t}$ is derived by using feature sets $\mathbf{f}_B(t - 1)$ and $\mathbf{f}_B(t)$. Instead of using only two video frames for estimating homography, we use the traces of all features in the consecutive frames. In Figure 6, we know that $f_B(x, y, t - 1)$ in frame $g(t - 1)$ moves to $f_B(x, y, t)$ in frame $g(t)$. However, transformation of $f_B(x, y, t - 1)$ by $M_{t-1,t}$ falls on $f'_B(x, y, t)$, because $M_{t-1,t}$ is derived by the least squares error minimization. Therefore, there is always a registration error between $f_B(x, y, t)$ and $f'_B(x, y, t)$. Similarly suppose $f_B(x, y, t + 1)$ is transformed to $g(t)$ such that $M^{-1}_{t,t+1}f_B(x, y, t + 1)$. Then the transformed point $f'_B(x, y, t + 1)$ also yields a registration error with respect to $f'_B(x, y, t)$. In an ideal case, all transformations

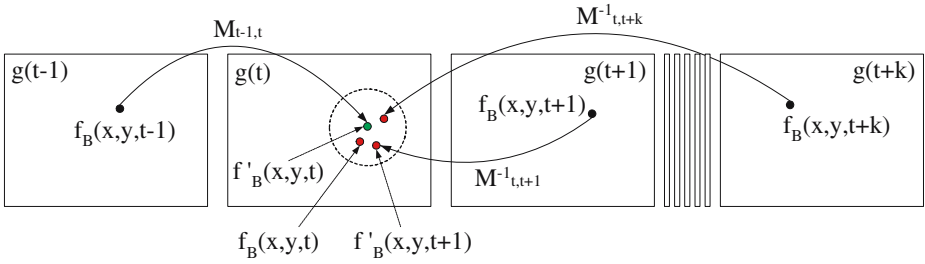


Fig. 6. Registration refinement using global features

of the same tracking features must yields the same coordinates. Therefore, we refine the pair-wise registration matrix $M_{t-1,t}$ by minimizing global registration error $\epsilon_{t-1,t}$ such that

$$\hat{f}_B(x, y, t) = \frac{1}{K} \sum_{k=0}^{K-1} M_{t,t+k}^{-1} f_B(x, y, t+k), \tag{11}$$

$$\epsilon_{t-1,t} = \sum_{\forall f_B \in \mathbf{f}_B(t-1)} \|f'_B(x, y, t) - \hat{f}_B(x, y, t)\|, \tag{12}$$

where $f'_B(x, y, t) = M_{t-1,t} f_B(x, y, t-1)$.

In the above equations, K is the maximum frame number before $f_B(x, y, t-1)$ fails to track. We also use the least-squares error minimization to iteratively minimize the global error criterion $\epsilon_{t-1,t}$. For every pair of feature sets $\{f'_B(x, y, t)\}$ and $\{\hat{f}_B(x, y, t+k)\}$, their centroids are moved to $(0, 0)$ and the average distance to their centroids is scaled to $\sqrt{2}$.

For each iteration, the refinement consists of two steps. In the first step, all pair-wise transforms from the first frame to the last frame are computed. In the second step, they are updated. When updating the transformations, we need to consider another error accumulation effect. Because we use the trace of a feature from the current to the last frame, there is no trace in the last frame. In other words, the last frame is the reference in the global registration. This causes another accumulation error which propagates from the last to the first frame. Let us show an example of accumulation effect in Figure 7. A transformation $M'_{t-1,t}$ is the updated form of $M_{t-1,t}$, similarly $M'_{t,t+1}$ is the updated form of $M_{t,t+1}$. Suppose we compute a transformation from $g(t-1)$ to $g(t+1)$, then $M'_{t,t+1}M'_{t-1,t}$ brings the position of a background feature to a green-colored (light gray) point in $g(t+1)$, while the correct position of the feature is the red-colored (dark gray) point. This is because the updated transformations have been derived by using the old coordinates of features in the first step. Therefore, we have to compensate the updated transformation to eliminate the error. Before multiplying the new transform from $g(t)$ to $g(t+1)$, ($M'_{t,t+1}$ in this example),

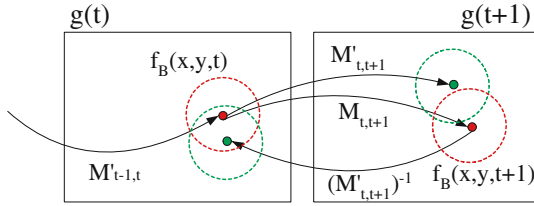


Fig. 7. Compensation of accumulation error due to refinement

we modify $M'_{t-1,t}$ so that $f_B(x, y, t - 1)$ transforms to the green-colored points instead of the red-colored point as follows:

$$M'_{t-1,t} \leftarrow (M'_{t,t+1})^{-1} M_{t,t+1} M'_{t-1,t}. \tag{13}$$

The two refinement steps are repeated until global registration error converges close to zero.

After refining the transformations, we fill the background images of removed object regions in all frames and generate a new video sequence. An image blending technique is also used to minimize the brightness blocking effect along the boundary of the object area.

4 Experimental Results

To obtain a video sequence of a moving object, we use a Canon GL-2 video camcorder. To minimize interlaced-scan noise, the video is recorded in progressive scan. From the original video sequence, we obtain a short video sequence which length is about 150-frame with 720×480 image resolution. Each frame is then saved to a TIFF or PPM image. After object removal and background completion, new frames are combined to generate a new video sequence. After defining the object window $W_O(0, x, y)$ in the first frame, corner features are extracted in a 300×300 image region, placing the object at its center.

Figure 8 shows results of the first experiment which uses a video sequence of a moving car. The sequence consists of 150 frames. The figure shows the original and new frames of number 0, 80, and 149. Figure 9 shows results of the second experiment. This sequence also consists of 150 frames. In each figure, an object-removed region is shown in detail. Background images are very accurately filled in the sequence. Table 1 shows registration error and processing time for the two objects. The registration error is measured at frame 0. Figure 10 shows panorama images of before and after refinement of the 'person' sequence. Accumulation of all frames shows the proposed global egistration yields very accurate results. Figure 11 plots the average of absolute registration error at some frames.



Fig. 8. Results of 'car' sequence. Top: Original frames. Middle: New frames. Bottom: Detail of removed region.



Fig. 9. Results of 'person' sequence. Top: Original frames. Middle: New frames. Bottom: Detail of removed region.



Fig. 10. Comparison of video mosaics. Top: Before refinement. Bottom: After refinement.

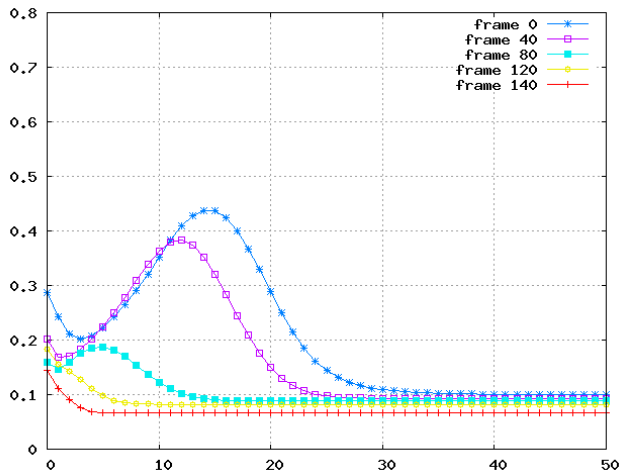


Fig. 11. Registration error (absolute average) of the 'person' sequence

Table 1. Registration error and processing time

Object	car person		
	Absolute error (pixel)	Average 0.114	0.105
	Max.	1.540 1.275	
Processing time (sec)	BT	42	42
	GR	36	28
	BC	4.5	5.2

BT: background tracking, GR: global registration (50 iterations), BC: background completion

5 Conclusions

We present a computer vision technique to remove a moving object in a video sequence. A new video sequence is generated by accurately filling the empty object area with its background image. To track the moving object, background features are tracked frame by frame using the KLT feature tracker. A global feature registration technique is introduced to refine all pair-wise transformations. Experimental results show that the proposed technique produces seamless background video sequences.

References

- [AG] A. Agarwala, A. Hertzmann, D. Salesin, and S. Seitz. "Keyframe-Based Tracking for Rotoscoping and Animation," *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2004)*, 2004.
- [BH] K. Bhat, M. Sapharishi, and P. Khosla, "Motion Detection and Segmentation Using Image Mosaics," *IEEE International Conference on Multimedia and Expo*, vol. 3, pp. 1577-1580, July, 2000.
- [SA] P. Sand and S. Teller, "Video Matching," *ACM Transactions on Graphics*, vol. 22, no. 3, pp.592-599, July 2004.
- [SU] Y. Sugaya and K. Kanatani, "Extracting moving objects from a moving camera video sequence," *Proceedings of the 10th Symposium on Sensing via Imaging Information*, pp. 279-284, June 2004.
- [TI] Y. Tian and A. Hampapur, "Robust, Salient Motion Detection with Complex Background for Real-time Video Surveillance," *IEEE Workshop on Applications on Computer Vision*, Breckenridge, Colorado, January 5-7, 2005.
- [WE] Y. Wexler, E. Shechtman, M. Irani, "Space-Time Video Completion," *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'04)*, vol 1, pp. 120-127, July 2004.
- [YA] A. Yamashita, T. Harada, T. Kaneko and K. Miura, "Removal of Adherent Noises from Images of Dynamic Scenes by Using a Pan-Tilt Camera," *Proceedings of the 2004 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2004)*, pp.437-442, Sendai (Japan), September 2004
- [ZH] Y. Zhang, J. Xiao, M. Shah, "Motion Layer Based Object Removal in Videos," *IEEE Workshop on Application on Computer Vision*, Jan 5-6, Breckenridge, Colorado, 2005.

Object Tracking Using Discriminative Feature Selection

Bogdan Kwolek

Rzeszów University of Technology,
W. Pola 2, 35-959 Rzeszów, Poland
bkwolek@prz.rzeszow.pl

Abstract. This paper presents an approach for evaluating multiple color histograms during object tracking. The method adaptively selects histograms that well distinguish foreground from background. The variance ratio is utilized to measure the separability of object and background and to extract top-ranked discriminative histograms. Experimental results demonstrate how this method adapts to changing appearances of both object undergoing tracking and surrounding background. The advantages and limitations of the particle filter with embedded mechanism of histogram selection are demonstrated in comparisons with the standard CamShift tracker and a combination of CamShift with histogram selection.

1 Introduction

This work addresses the issue of on-line selection of discriminative color features during object tracking. Feature selection is a process of mapping the original data into more effective features [1]. If features with little discrimination capabilities are selected, even a good algorithm can lead to poor tracking performance. On the other side, if discriminative features are selected the tracking system can be simplified and thus a limited number of CPU cycles can be sufficient. The most tracking methods operate using only a fixed set of features that are determined in advance. As stated in [2][3], comparatively little work has been done in building tracking systems, which can select most discriminative features on-line. In their work [4], Shi and Tomasi have pointed out that discriminative features are just as equally important as good tracking algorithms.

Selecting a low-dimensional discriminative feature set can improve tracker performance. The goal of dimensionality reduction is to preserve most of the relevant information of the original data according to some optimality criteria. Methods such as principal component analysis (PCA), independent component analysis (ICA) and linear discriminant analysis (LDA) are exemplars of algorithms finding a mapping between the original feature space and a lower dimensional feature space [5]. These methods involve feature transformation and create a set of transformed features rather than a subset of the original features. In work [3] feature extraction is achieved by PCA and the number of dimensions is determined by the pre-defined proportion of eigenvalues. Weights are assigned

to each pixel and the mean-shift algorithm [1][6] is utilized to perform tracking. The variance ratio is employed to evaluate the degree of the saliency for the foreground in the likelihood image. The main limitation of this approach is that some visual information from the original image can be lost by the projection. The work [2] also uses the likelihood image to combine feature spaces and to select better ones. A method for evaluating several feature spaces while the tracking process proceeds is proposed. It selects the best feature space among candidates that are constructed by different linear combinations of the three color channels from the RGB color space. The method utilizes the previous frame as the training frame to perform a feature selection and then utilizes the current frame as the test frame for foreground-background classification. The features are ranked on the basis of a variance test for the distinctiveness between object and background. Improved tracking performance to standard mean-shift based tracking algorithm has been reported. However, the creation of 49 likelihood images is time consuming.

The importance of the background appearance for tracking has been emphasized in other work [7]. This algorithm maintains a pool of discriminant functions each distinguishing an object pattern against the background patterns that are currently relevant. A searching for the region that best matches the targets and simultaneously avoids background patterns seen previously is embedded in this algorithm. Combining both labeled and unlabeled data is utilized in discriminant expectation maximization (D-EM) algorithm [8] to automatically select a good color space. The basic idea of D-EM is to identify some similar samples in the unlabeled set to grow the labeled data set and then to apply a supervised technique on such enlarged labeled set. Both background and foreground are represented by mixtures of Gaussians.

In work [9] a dynamic switching between five predetermined color spaces takes place in order to improve the performance of face tracking. The selection of color space is done using the ratio of flesh probability pixels within the internal and external face windows with concentric location.

Traditional appearance based representations construct appearance models from examples in training data sets and then utilize such models to track the object of interest. Color histograms [10] that are invariant to some degree of viewpoint change are often used to construct appearance models. Appearance based representations can be very useful in construction of fast and effective tracking systems [11][12][6][13]. For example, the scale invariant feature transform (SIFT) [14] employs a histogram of gradient that is scale and rotation invariant.

Recent work on on-line selection of discriminative features for tracking as well as the success of appearance methods in tracking inspired us to base our tracking method on color histograms. We employ a selection algorithm that maintains a pool of histograms to select histograms yielding more discriminative power. A pool of histograms assigning the various number of bins to each of the color component of the utilized color space is maintained. Our contribution to on-line

selection of discriminative features is a method which allows to select the most appropriate color histograms in the current context.

The rest of the paper is organized as follows. The next Section contains a description of evaluating feature discriminability. Section 3. is devoted to object tracking. In Section 4. we outline CamShift based tracking with feature selection. In Section 5. we present all ingredients of our probabilistic tracker with adaptive feature selection and report results which were obtained in experiments. Finally, some conclusions follow in the last Section.

2 Evaluating Feature Discriminability

At the beginning of this section, we show how the log likelihood ratios are computed. The feature space will be presented as the second topic. A description of feature discriminability ends this section.

2.1 Likelihood Ratios of Foreground and Background Histograms

A variety of parametric and non-parametric statistical methods can be utilized to represent color distributions of homogeneous colored areas. The histogram is the oldest and most widely applied non-parametric density estimator. It is computed by counting the number of pixels in a region of interest that have given color. The colors are quantized into bins. This operation allows similar color values to be clustered as single bin. By normalizing the histogram by the number of elements in it we form the discrete probability density representing the given object. Methods using histograms techniques are effective only when the number of bins can be kept relatively low and when sufficient data amounts are in disposal. Histogram based methods are only suitable for low dimensional data spaces because as the number of dimensions expand, the number of bins should grow exponentially.

Given a foreground histogram and a background histogram, the log-likelihood ratio for a pixel with color \mathbf{u} is given by [3]:

$$L(\mathbf{u}) = \log \frac{\max(p(\mathbf{u}), \delta)}{\max(q(\mathbf{u}), \delta)}, \quad (1)$$

where δ is a very small number, whereas $p(\mathbf{u})$, $q(\mathbf{u})$ represent the discrete probability density of color pixels in the foreground and background, respectively. Colors that are shared by both foreground and background have values $L(\mathbf{u})$ which tend towards zero. The likelihood image can be computed by back-projecting the ratio for each pixel in the image. Then the salient region in object of interest can be identified by pixels with high likelihood ratios. Such regions, extracted on the basis of different features can be employed to extract a binary mask identifying the object.

2.2 Feature Space

The color histograms are usually extracted through assigning to each color channel a fixed number of bits, determined a priori. Such approaches ignore the fact that both foreground and background appearance undergo changes as the target moves. The ability to distinguish between object and background can be insufficient when histograms assigning each color channel a fixed number of bits have been chosen. A color histogram with specific combination of bins for each color channel and possessing good discrimination capabilities for tracking a car in front of green background can perform poorly when colors in the background change their values.

In our approach we maintain identical number of total bins in all candidate histograms. The set of candidate histograms is composed of linear combinations of bin numbers assigned to color channels. In our implementation the RGB color space is utilized and the number of histogram bins m is set to 512. With this histogram length and assuming that the number of bins for each color channel can take the values 2^b , where $b = 0, 1, \dots, 5$, we can construct a pool of candidate histograms. Table 1. presents a set of candidate histograms that was utilized in this work. Given a pixel at position x_i , the bin index of 1D histogram is computed as follows:

$$\text{idx} = c_R(x_i) + c_G(x_i) * m_R + c_B(x_i) * m_R * m_G \quad (2)$$

where the function $c_j(x) : \mathcal{R}^2 \rightarrow \{1, \dots, m_j\}$ associates the value of pixel at location x_i to bin number, $j \in \{R, G, B\}$, whereas R, G, B denote color channels.

Table 1. Number of bins assigned to each color channel in the set of candidate histograms

	0	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15
m_R	8	4	4	8	16	16	8	2	16	16	1	1	16	32	16	32
m_G	8	16	8	4	4	8	16	16	2	16	16	32	32	16	1	1
m_B	8	8	16	16	8	4	4	16	16	2	32	16	1	1	32	16

2.3 Feature Discriminability

The foreground and background pixels are sampled using center-surround approach in which an internal rectangle covers the object, while a larger surrounding rectangle represents the background. Following the suggestion in [2], the grade of the salience for the foreground and the likelihood image can be expressed by the variance ratio:

$$\text{VR}(L; p, q) = \frac{\text{var}(L; (p+q)/2)}{\text{var}(L; p) + \text{var}(L; q)} \quad (3)$$

where $\text{var}(L; a) = \sum_i a(i)L^2(i) - [\sum_i a(i)L(i)]^2$. The log likelihood images associated with features of high variance ratio correspond to good features in terms

of foreground and background separability. On the basis of the variance ratio we extract top-ranked discriminative histograms.

3 Object Tracking

There are, generally, two types of tracking algorithms: deterministic and probabilistic. The mean-shift algorithm and CamShift are the most famous deterministic tracking algorithms. They may be trapped in local minima and generally can not recover from temporary failure. This problem can ameliorate probabilistic methods built on particle filters. They achieve robustness to clutter and occlusion by maintaining multiple hypotheses over the state space. At the beginning of this section we describe the CamShift algorithm. The second part of this section is devoted to particle filtering.

3.1 CamShift

CamShift tracking algorithm is based on a robust non-parametric technique called mean-shift to seek the nearest mode of probability distribution. The searching starts from the final location in the previous frame and proceeds iteratively to find the nearest mode. The value of each pixel in the probability image represents the probability that the pixel belongs to the object of interest. The object probability density image $P(x, y)$ is extracted through thresholding the log likelihood image.

The mean location of the distribution within the search window is computed using moments [15][12]. It is given by:

$$x_1 = \frac{\sum_x \sum_y xP(x, y)}{\sum_x \sum_y P(x, y)}, \quad y_1 = \frac{\sum_x \sum_y yP(x, y)}{\sum_x \sum_y P(x, y)} \quad (4)$$

where x, y range over the search window. The eigenvalues (major length and width) of the probability distribution are calculated as follows [15][12]:

$$l = 0.707\sqrt{(a+c) + \sqrt{b^2 + (a-c)^2}}, \quad w = 0.707\sqrt{(a+c) - \sqrt{b^2 + (a-c)^2}} \quad (5)$$

where

$$a = \frac{M_{20}}{M_{00}} - x_1^2, \quad b = 2\frac{M_{11}}{M_{00}} - x_1y_1, \quad c = \frac{M_{02}}{M_{00}} - y_1^2, \quad M_{00} = \sum_x \sum_y P(x, y), \\ M_{20} = \sum_x \sum_y x^2P(x, y), \quad M_{02} = \sum_x \sum_y y^2P(x, y).$$

The algorithm repeats the computation of the centroid and repositioning of the search window until the position difference converges to some predefined value, that is, changes less than some assumed value. Relying on the zero-th moment M_{00} the CamShift adjusts the size of the search window in the course of its operation. It requires the selection of the initial location and size of the search window. The algorithm outputs the position, dimensions, and orientation of object undergoing tracking. It can be summarized in the following steps [12]:

1. Set the search window at the initial location (x_0, y_0) .
2. Determine the mean location in the search window (x_1, y_1) .
3. Center the search window at the mean location computed in Step 2, set the window size to zero-th moment M_{00} .
4. Repeat Steps 2 and 3 until convergence.

3.2 Particle Filtering

The effectiveness of object tracking in image sequences has been greatly improved with the development of particle filtering. The particle filter is an algorithm for estimating the posterior state of a dynamic system over time where the state cannot be measured directly, but may be estimated at the current time-step t . Particle filters are attractive for nonlinear models, multi-modal, non-Gaussian or any combination of these models for several reasons. They utilize imperfect observation and motion models and incorporate noisy collection of observations through Bayes rule. The ability to represent multimodal posterior densities allows them to globally localize as well as relocalize the object of interest in case of temporal failure during tracking. Particle filters are any-time because by supervising the number of samples on-line they can adapt to the available computational resources.

Two important components of each particle filter are motion model $p(\mathbf{x}_t | \mathbf{x}_{t-1})$ describing the state propagation and observation model $p(\mathbf{z}_t | \mathbf{x}_t)$ describing the likelihood that a state \mathbf{x}_t causes the observation \mathbf{z}_t . Starting with a weighted particle set $S = \{(\mathbf{x}_{t-1}^{(n)}, \pi_{t-1}^{(n)}) | n = 1 \dots N\}$ approximately distributed according to $p(\mathbf{x}_{t-1} | \mathbf{z}_{1:t-1})$ the particle filter operates through predicting new particles from a proposal distribution. To give a particle representation $S = \{(\mathbf{x}_t^{(n)}, \pi_t^{(n)}) | n = 1 \dots N\}$ of the posterior density $p(\mathbf{x}_t | \mathbf{z}_{1:t})$ the weights of particles are set to $\pi_t^{(n)} \propto \pi_{t-1}^{(n)} p(\mathbf{z}_t | \mathbf{x}_t^{(n)}) p(\mathbf{x}_t^{(n)} | \mathbf{x}_{t-1}^{(n)}) / q(\mathbf{x}_t^{(n)} | \mathbf{x}_{t-1}^{(n)}, \mathbf{z}_t)$. When the proposal distribution from which particles are drawn is chosen as the distribution conditional on the particle state at the previous time step, the importance function reduces to $q(\mathbf{x}_t^{(n)} | \mathbf{x}_{t-1}^{(n)}, \mathbf{z}_t) = p(\mathbf{x}_t^{(n)} | \mathbf{x}_{t-1}^{(n)})$ and the weighting function takes the form $\pi_t^{(n)} \propto p(\mathbf{z}_t | \mathbf{x}_t^{(n)})$. This simplification leads to a variant of a particle filter, CONDENSATION [16]. From time to time the particles should be resampled according to their weights to avoid degeneracy [17].

4 CamShift Based Tracking with Feature Selection

The tracking algorithm we present here follows the idea of selection of discriminative features on-line, which is presented in [2]. In this section we examine a selection algorithm to determine how well each histogram distinguishes object from background in the current frame. The feature selection algorithm is embedded in CamShift based tracking system.

The CamShift algorithm is utilized to find the estimate of the 2D object location of the object in the frame. Using the estimated object location as well

as an object mask we extract all candidate histograms. Afterwards, we select the top-ranked discriminative histograms on the basis of the variance ratio. The best three histograms are used to extract the likelihood images for the next frame. Using such likelihood images we extract the compound likelihood image, which is a simple weighted average. After thresholding the compound image we get the binary image. The compound image is subjected to CamShift.

The algorithm iterates through frames and chooses new sets of discriminative histograms. All candidate histograms representing both background and foreground are adapted over time. To avoid model drift the histograms are adapted using linear combination of current observed histograms, the histograms from the last frame as well as histograms from the first frame. The accommodation coefficients were determined experimentally under assumption that the object appearance will not change drastically over the tracking sequence. The set of features used for tracking changes while the tracking process proceeds. Figure 1. depicts some probability images corresponding to the best and worst pair of foreground and background color histograms, in terms of foreground and background separability.

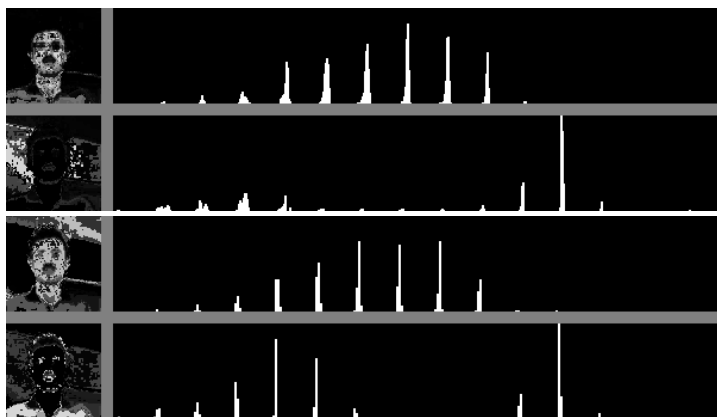


Fig. 1. Probability images of foreground/background and corresponding histograms (frame #50 in the sequence of images demonstrated in Fig. 2). The probability images and histograms for the most discriminative feature are in upper row. The images and histograms for least discriminative feature are in bottom row.

The images from middle row of Fig. 2. illustrate the failure of standard CamShift algorithm. The standard CamShift algorithms operate using only a fixed set of three histograms and do not change this pre-selected set while the tracking process proceeds. During tracking in varying illumination conditions the tracker is affected by similar background color, leading to tracking failure.

The tracker with histogram selection detects which colors in the model are similar to colors in background and tries to choose the histograms that allow for better foreground/background separation. This property can be observed in

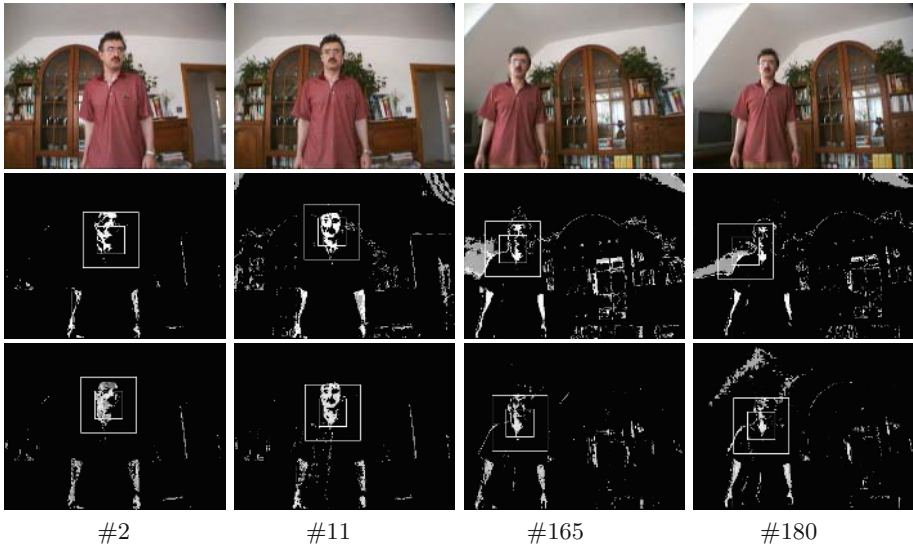


Fig. 2. Face tracking in varying illumination conditions using CamShift. Raw color images (top row), object probability images-no feature selection takes place (middle row), object probability images-feature selection (bottom row).

Fig. 3. We see that during tracking under illumination changes, frame #165 in Fig. 2., the tracker adapts to changing appearances of both tracked object and the background. Our algorithm continues the tracking whereas the standard CamShift with pre-selected histogram pool suddenly loses the object.

Figure 3. shows how the selection of the best histogram in sequence of images from Fig. 2. evolves over time. For most frames of the sequence the algorithm selects the histogram number zero, see Tab. 1., which assigns the equal number of bins to all color channels. In several frames the algorithm selects thirteen pair of histograms. The selection mechanism supports the tracking and allows the object model to adapt to current conditions and background distractions.

5 Probabilistic Tracking with Feature Selection

In our approach we consider only the location $\mathbf{d} = (x, y)$ in the image coordinate system, the window scale s and the histogram number as the state variables to be estimated. One way to model the transition of the state is using a random walk which can be described by

$$\mathbf{x}_{t+1} = \mathbf{x}_t + \eta. \quad (6)$$

A Gaussian noise $N(0, \nu^2)$, where ν^2 is typically learned from training sequences, has been added to the first three state variables, whereas the evolution of the histogram number in such a hybrid state particle filter was modeled using a

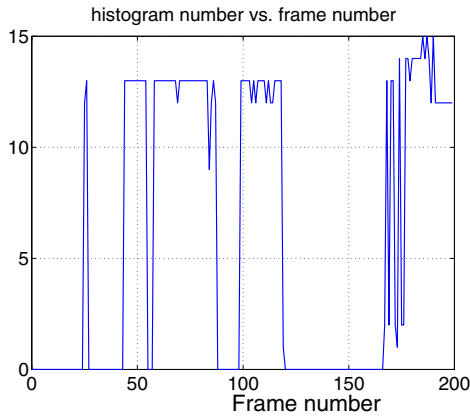


Fig. 3. Number of the best histogram for tracking versus frame number

probability distribution over possible histogram numbers. Such a choice was motivated by observation that the frame to frame position differences in our test sequences are not too large.

The observation model must favor candidate object locations close to the true object locations as well as favor histograms yielding better separability between foreground and background. We therefore need to consider the object probability in the object window given the state of the particle. An iterative mode-seeking in the form of the mean-shift algorithm can be applied to shift the particles to high weight areas [18][19].

The kernel based methods of density estimation construct an estimate of the true density distribution through placing a kernel function on each sample. The estimate of the posterior distribution $p(\mathbf{x}_t | \mathbf{z}_t)$ with kernel K can be formulated as follows:

$$\hat{p}(\mathbf{x}_t | \mathbf{z}_t) = \sum_{n=1}^N K_h(\mathbf{x}_t - \mathbf{s}_t^{(n)}) \pi_t^{(n)} \tag{7}$$

where $K_h(\mathbf{x}_t - \mathbf{s}_t^{(n)}) = \frac{1}{N h^d} K(\frac{\mathbf{x}_t - \mathbf{s}_t^{(n)}}{h})$, and h is the kernel bandwidth. For the radially symmetric kernel we have $K(\mathbf{x}_t - \mathbf{s}_t^{(n)}) = c k(\|\mathbf{x}_t - \mathbf{s}_t^{(n)}\|)$, where c is a normalization constant which makes the integral $K(\mathbf{x}_t - \mathbf{s}_t^{(n)})$ to one, and $k(r) = k(\|\mathbf{x}_t - \mathbf{s}_t^{(n)}\|)$ is called the profile of the kernel K . In our particle filter we employ the Epanechnikov kernel that is defined as:

$$K_E(\mathbf{x}) = \begin{cases} \frac{1}{2} c_d^{-1} (d + 2) (1 - \|\mathbf{x}\|^2) & 0 \leq \|\mathbf{x}\| \leq 1 \\ 0 & \|\mathbf{x}\| > 1 \end{cases} \tag{8}$$

Given a particle set and the associated weights $\{\pi_t^{(n)}\}_{n=1}^N$, the particle mean is determined by

$$m(\mathbf{s}_t^{(n)}) = \frac{\sum_{i=1}^N H_h(\mathbf{s}_t^{(n)} - \mathbf{s}_t^{(i)}) \pi_t^{(i)} \mathbf{s}_t^{(i)}}{\sum_{i=1}^N H_h(\mathbf{s}_t^{(n)} - \mathbf{s}_t^{(i)}) \pi_t^{(i)}}, \tag{9}$$

where $h(r) = -k'(r)$ is in turn a profile of kernel H_h . It can be shown that the mean-shift vector $m(\mathbf{x}) - \mathbf{x}$ always points toward steepest ascent direction of the density function.

The choice of bandwidth h is of crucial importance in kernel based density estimation. A small value can generate a very ragged density approximation with many peaks, while a large value of h can produce over-smoothed density estimates. In particular, if the bandwidth of the kernel is too large, significant features of the distribution, like multi-modality can be missed.

The mode-seeking continues searching until a maximum number of iterations has been reached or until the Euclidean distance between the corresponding modes in the last two iterations is below an empirically determined threshold. We scale down the kernel bandwidth at each mean-shift iteration in order to concentrate on the most dominant modes. Following mode-seeking, the most dominant mode is extracted on the basis of weighted average over all particles within the kernel. The tracking scheme can be summarized as follows: $p(\mathbf{x}_{t-1} | \mathbf{z}_{t-1}) \xrightarrow{\text{dynamics}} p(\mathbf{x}_t | \mathbf{z}_{t-1}) \xrightarrow{\text{measurement}} p(\mathbf{x}_t | \mathbf{z}_t) \xrightarrow{\text{mean-shift}} \hat{p}(\mathbf{x}_t | \mathbf{z}_t)$. Each particle can only change its location during mean-shift iterations. The following observation model is utilized:

$$p(\mathbf{z}_t | \mathbf{x}_t) = (1.0 - \exp(-\lambda_1 \text{VR}^2)) \times (1.0 - \exp(-\lambda_2 \text{Pr}^2)) \tag{10}$$

where VR denotes the variance ratio and Pr is the mean probability in the object window.



Fig. 4. The results of tracking using CamShift (top row) and particle filter (bottom row)

To test our probabilistic tracker we performed experiments using various test sequences. Experimental results that are depicted in Fig. 4. indicate that due to its Monte Carlo nature, the particle filter better handles confusions that are caused by similar colors in the background. Both CamShift and probabilistic tracker were initialized with a manually selected object region of size 20x20 in frame #2799.

The algorithms were implemented in C/C++ and run on a 2.4 GHz PIV PC. The average number of mean-shift iterations per frame is 2.9. The tracker runs with 60 particles at frame rates of 12-13 Hz. All experiments were conducted on images of size 320x240.

6 Conclusions

We have presented an approach for evaluating multiple color histograms during object tracking. The elaborated method adaptively selects histograms that well distinguish foreground from background. It employs the variance ratio to quantify the separability of object and background and to extract top-ranked discriminative histograms. The superiority of CamShift based tracker using the histogram selection over the traditional CamShift tracking arises because the variance ratio when applied to log likelihood images, which are computed on the basis of various candidate histograms, yield very useful information. Our algorithm evaluates all candidate histograms to determine which ones provide better separability between foreground and background. By employing the histogram selection, the modified CamShift can track objects in case of dynamic background. The particle filter with the embedded selection of histograms is able to track objects reliably during varying lighting conditions. To show advantages of our approach we have conducted several experiments on real video sequences. Currently, only RGB space is used. The performance of the visual tracker could be much better if other color spaces such as HSI could be utilized within this tracking framework.

Acknowledgment

This work has been supported by MNSzW within the project 3 T11C 057 30.

References

1. Fukunaga, K.: Introduction to Statistical Pattern Recognition. 2 edn. Acad. Press (1990)
2. Collins, R., Liu, Y.: On-line selection of discriminative tracking features. In: Proc. of the Int. Conf. on Computer Vision, Nice, France (2003) 346–352
3. Han, B., Davis, L.: Object tracking by adaptive feature extraction. In: Proc. of Int. Conf. on Image Processing (ICIP), Singapore (2004) III:1501–1504
4. Shi, J., Tomasi, C.: Good features to track. In: Proc. of IEEE Conf. on Computer Vision and Pattern Recognition, Seattle, Washington (1994) 593–600

5. Duda, R.O., Hart, P.E., Stork, D.G.: *Pattern Classification*. 2 edn. John Wiley & Sons, Inc. (2001)
6. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: *Proc. of IEEE Conf. on Comp. Vision and Pattern Recognition*, Hilton Head, SC (2000) 142–149
7. Nguyen, H.T., Smeulders, A.: Tracking aspects of the foreground against the background. In: *8th European Conf. on Computer Vision*, Prague, Czech Republic (2004) 446–456
8. Wu, Y., Huang, T.S.: Color tracking by transductive learning. In: *Proc. of IEEE Conf. on Computer Vision and Pattern Recognition*, Hilton Head Island, South Carolina (2000) 133–138
9. Stern, H., Efros, B.: Adaptive color space switching for tracking under varying illumination. *Image and Vision Computing* **23** (2005) 353–364
10. Swain, M.J., Ballard, D.H.: Color indexing. *Int. Journal of Computer Vision* **7** (1991) 11–32
11. Birchfield, S.: Elliptical head tracking using intensity gradients and color histograms. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, Santa Barbara, CA (1998) 232–237
12. Bradski, G.R.: Computer vision face tracking as a component of a perceptual user interface. In: *Proc. IEEE Workshop on Applications of Comp. Vision*, Princeton (1998) 214–219
13. Perez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-based probabilistic tracking. In: *7th European Conf. on Computer Vision*, Copenhagen, Denmark (2002) 661–675
14. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *Int. J. of Computer Vision* **60** (2004) 91–110
15. Horn, B.K.P.: *Robot Vision*. The MIT Press (1986)
16. Isard, M., Blake, A.: Contour tracking by stochastic propagation of conditional density. In: *4th European Conf. on Computer Vision*, Cambridge, UK (1996) 343–356
17. Doucet, A., Godsill, S., Andrieu, C.: On sequential monte carlo sampling methods for bayesian filtering. *Statistics and Computing* **10** (2000) 197–208
18. Han, B., Zhu, Y., Davis, L.: Incremental density approximation and kernel-based bayesian filtering for object tracking. In: *Int. Conf. on Computer Vision and Pattern Recognition*, Washington, DC (2004) 638–644
19. Fritsch, J., J., Kwolek, B.: Kernel particle filter for real-time 3d body tracking in monocular color images. In: *IEEE Int. Conf. on Face and Gesture Recognition*, Southampton, UK, IEEE Computer Society Press (2006) 564–567

Color-Based Multiple Agent Tracking for Wireless Image Sensor Networks

Emre Oto, Frances Lau, and Hamid Aghajan

Wireless Sensor Networks Lab
Department of Electrical Engineering
Stanford University, Stanford, CA 94305
{eoto, flau, aghajan}@Stanford.edu

Abstract. This paper presents an implementation of a color-based multiple agent tracking algorithm targeted for wireless image sensor networks. The proposed technique is based on employing lightweight algorithms and low-bandwidth data communication between multiple network nodes to track the path of autonomous agents moving across the fields of view (FOV) of the sensors. Segmentation techniques are applied to find the agents within the FOV, and a color histogram is constructed using the hue values of the pixels corresponding to agents. This histogram is used as a means of identification within the network. As such, the algorithm is able to reliably track multi-colored agents of irregular shapes and sizes and can resolve identities after collisions. The proposed algorithm has low computational requirements and its complexity scales linearly with the size of the network, so it is feasible in low-power, large-scale wireless sensor networks.

1 Introduction

The problem of tracking people, cars and other moving objects has long been at the focal point of many technical disciplines. Many techniques have been proposed and implemented to track multiple objects with multiple cameras, most of which employ stochastic models such as Kalman filtering, particle filtering, and condensation algorithms to overcome problems of occlusion, noisy observations and other visual artifacts [1]. Nguyen et al. [2] have implemented a distributed tracking system employing a Kalman filter to track multiple people within a room monitored by multiple cameras with overlapping fields of view. A similar solution to the same problem has been proposed by Chang et al. [3], who use a Bayesian network and Kalman filtering to establish correspondence of subjects between subsequent frames. There have also been applications of multiple object tracking algorithms implemented in sensor networks, such as the method demonstrated by Chang and Huang [4], in which distributed processing between multiple trackers is employed and data fusion is realized by an enhanced Kalman filter.

Color-based tracking methods that combine color information with statistical methods have also been in practice. Liu et al. [5] have implemented an algorithm based on particle filtering and color histogram information for object tracking. A

similar color-based Kalman particle filter algorithm has been implemented for the same application by Limin [6]. Perez et al. [7] have used a hue-saturation histogram with a particle filter based probabilistic technique for tracking in cluttered environments. One common aspect of all the mentioned color-based algorithms is that they have been designed for use outside of the wireless sensor networks domain, and require significant processing load by a centralized processing unit.

In this paper, we propose a hue histogram based multiple object, multiple camera tracking algorithm that is intended to be simple in nature to find potential use in wireless sensor nodes. The main contribution of this paper is to demonstrate that identity management of multi-colored agents of irregular shapes traveling on a cluttered background in an image sensor network can be performed using deterministic histogram matching techniques on the hue histogram, a metric that is small in size to be communicated between sensor nodes, and yet provides reasonable resilience against changes in illumination.

The Color-Based Multiple Agent Tracking (COBMAT) algorithm uses multiple image sensors to track the movement of autonomous mobile agents or targets traveling within overlapping or non-overlapping fields of view (FOVs) of overhead and side view cameras. The objective of the algorithm is to keep track of the agent positions in a distributed fashion, without the need for a centralized control scheme. The algorithm achieves this by communicating agent information between sensors and by performing identity management using hue histograms. The algorithm does not predict the path of agents traveling between non-overlapping FOVs. This is coherent with practical applications where cameras may be used to monitor specific areas of a field, e.g. rooms of a building or particular sectors of a military field.

The image sensors are assumed to be localized by an algorithm such as in [8]. In some tracking applications, the objective may be to know which network node is tracking the agents or targets of interest. In other applications, the global coordinates of the agent at observation times may also be required. In the case of overhead cameras, the agent's global position in the area spanned by the network can be calculated by the tracking node and broadcasted to other nodes. In the case of side view cameras, exchange of image plane agent positions between the nodes that simultaneously observe the agent can result in determining the global position of the agent.

The COBMAT algorithm relies on a background subtraction and segmentation routine to obtain the blobs in each frame. The color histogram of each blob is extracted by calculating the hue of each pixel from the RGB (red, green, blue) values and then binning the hue values to create a histogram. This hue histogram is compared to those belonging to the agents identified in previous observations to associate new blobs with agents tracked in the previous FOV. The blobs that could not be identified as existing agents are compared against a local database called the "Potential Agents Database", which contains the histograms of the agents detected by neighboring sensors. Each node broadcasts every new agent entering its FOV to its one-hop neighbors and each sensor records these messages in its "Potential Agents Database". This is done to determine if the new agent has been previously identified by the network or if it is a new entry to the realm of the network.

2 System Overview

COBMAT is an algorithm intended for use with image sensor nodes in tracking agents or targets within the field of view (FOV) of one sensor and in providing inter-sensor broadcast of information on agents or targets traveling between the FOVs of different sensors. Fig. 1 illustrates a schematic for a network of image sensors covering an area of monitoring interest.

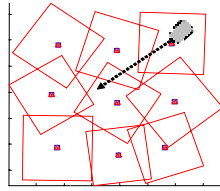


Fig. 1. Illustration of the network of overhead image sensors and a moving agent

The overall operation of the COBMAT algorithm can be epitomized as in Fig. 2. Within a single FOV, the blobs are extracted with the Blob Extraction Module, and then identified using the Identity Management Module. The Inter-sensor Communication Module handles the broadcast of new entries to a sensor's FOV to its one-hop neighbors. These three modules are described in the following sections. The distributed nature of COBMAT enables it to be scalable to large camera networks.

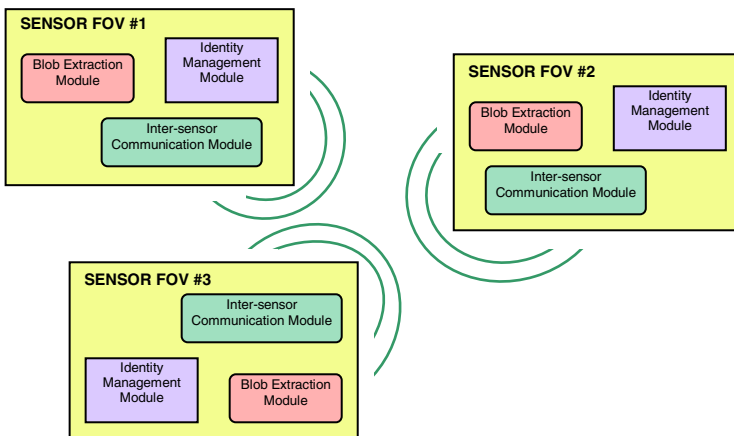


Fig. 2. Overall system block diagram of the COBMAT algorithm

2.1 Blob Extraction

Blob extraction is performed on the still images obtained at each frame instant. This operation relies on background subtraction followed by a segmentation routine. Background subtraction identifies the change mask by thresholding the difference

between the intensities of each of the color channels in the current frame and the background image. To be identified as a changed pixel, the pixel values in any of the red, green or blue channels of the pixel must have changed by more than a certain threshold. During experimental studies, this threshold was selected according to the illumination of the environment. In a practical system, however, the threshold must be adaptively determined as a function of the ambient illumination.

Due to the manner in which changed pixels are found, it is important that the background image manifest the current lighting conditions. Thus, background subtraction assumes that an up-to-date image of the sensor FOV devoid of agents is available. A measure that has been taken against illumination changes is color balancing of the background image and the frame prior to background subtraction. A color balancing algorithm based on the gray-world assumption has been observed to eliminate spurious blobs arising in the case of low illumination.

Following background subtraction, a small-region removal is run on the mask image to remove artifacts due to slight noise effects. This is followed by a large-region removal algorithm, applied on the inverted mask, to fill the holes within the regions identified as blobs. These regions are then labeled, and the mask resulting from the two region removal operations is applied on the original image to retrieve the blob pixel values. The hue of each pixel is then calculated and the hue histogram of each blob is produced. The hue histograms used in this study were of 36 bins.

The position of the blob, i.e. the target location within the FOV, is found as the median of the rows and the median of the columns comprising the blob.

2.2 Identity Management

In each frame instant, the hue histograms of the blobs are extracted, and these histograms are compared to those belonging to the agents in the previous frame to identify which agents have remained in the FOV, which agents have just entered, and which have left the FOV. The histogram matching routine employed at this stage relies on two different distance measures to decide whether the two histograms are the same: the Euclidean Distance (ED) and the Vector Cosine Distance (VCD) measures. The ED between two histograms $h[n]$ and $g[n]$ of length N is given as in (1).

$$ED = \sqrt{\sum_{n=1}^N (g[n] - h[n])^2} . \quad (1)$$

Treating the histograms $h[n]$ and $g[n]$ as two vectors in \mathbb{R}^N , ED is the norm of the difference vector $h[n]-g[n]$. VCD is a measure of proximity proposed by Sural et al. [9] that derives itself from the Euclidean geometry, and is the angle between the two vectors. This projection angle $\theta(g[n], h[n])$ can be calculated as given below:

$$\theta(g[n], h[n]) = \cos^{-1} \left(\frac{\sqrt{\sum_{n=1}^N (g[n] \cdot h[n])}}{|g| \cdot |h|} \right) , \quad (2)$$

$$\text{where } |g| = \sqrt{\sum_{n=1}^N (g[n])^2} \quad \text{and} \quad |h| = \sqrt{\sum_{n=1}^N (h[n])^2} .$$

Two agent histograms are inferred to be the same if both the ED and VCD tests yield that the histograms are the same, which is decided if the distance calculated by the two measures are individually smaller than the thresholds. Through experimentation, an ED of 1 and a separation angle of $\pi/4$ were found to be appropriate thresholds for correct detection even under varying illumination conditions.

It can easily be observed that both the ED and VCD are scalar norms calculated on a bin-by-bin basis. By their nature these norms fail even if the histograms are very similar in shape but are shifted with respect to one another. The effect of such a shift can be amplified in the case of the hue histogram since it is an angular measure wrapping around from 2π to 0 radians. To alleviate the bin shift and wrap around problems, the ED and VCD are calculated as follows:

$$ED_{g,h} = \min(ED(g[n],h[n]), ED(g[n],h[(n-1)]_N), ED(g[n],h[(n+1)]_N), ED(\text{avg}(g[n]),\text{avg}(h[n]))) \quad (3)$$

$$\theta_{g,h} = \min(\theta(g[n],h[n]), \theta(g[n],h[(n-1)]_N), \theta(g[n],h[(n+1)]_N), \theta(\text{avg}(g[n]),\text{avg}(h[n]))) , \quad (4)$$

where $(()_N$ represents a circular shift in modulo N , and the $\text{avg}()$ function is the application of an averaging window of size N_w . The circular shift by 1 covers single bin shifts of the histograms with respect to each other, while the window averaging leverages the result by smoothing when the two histograms are shifted by more than a single bin. In our empirical studies, we observed that $N_w = 3$ achieved sufficient smoothing.

Fig. 3 depicts the identity management algorithm employed in each sensor. The histograms of agents traveling in the FOV in the most recent frame are stored in cache memory for easy access when performing identity associations of agents in the current frame. At each frame instant identity association is performed by means of histogram matching as explained above. If a blob in the current frame can be matched to any of the agents, the position of that agent is updated in the cache.

If no match can be found in the current FOV, then the position of the blob is considered. If all the cameras are overhead cameras, then the background on which the agents travel is a 2-D plane, so new agents may not emerge from inside the FOV. Using this fact, we only consider the blobs that emerge near the edge of the FOV to be new agents, and assume all other blobs are artifacts caused by sudden lighting changes. The threshold distance used to decide if the agent is far away from the edge or not was determined by experimentation as it depends on the size of the FOVs and the agents as well as the frame rate and the range of possible agent speeds. In the situation where side-view cameras are used, this condition for the position of the blob is not used.

Agents that are new to a FOV are first sought in the ‘‘Potential Agents Database’’, which contains the histograms and labels of the agents communicated by the sensor’s one-hop neighbors upon entrance to their FOVs. If the agent is identified in the ‘‘Potential Agents Database’’, the matching entry is copied from the database to the cache. If the

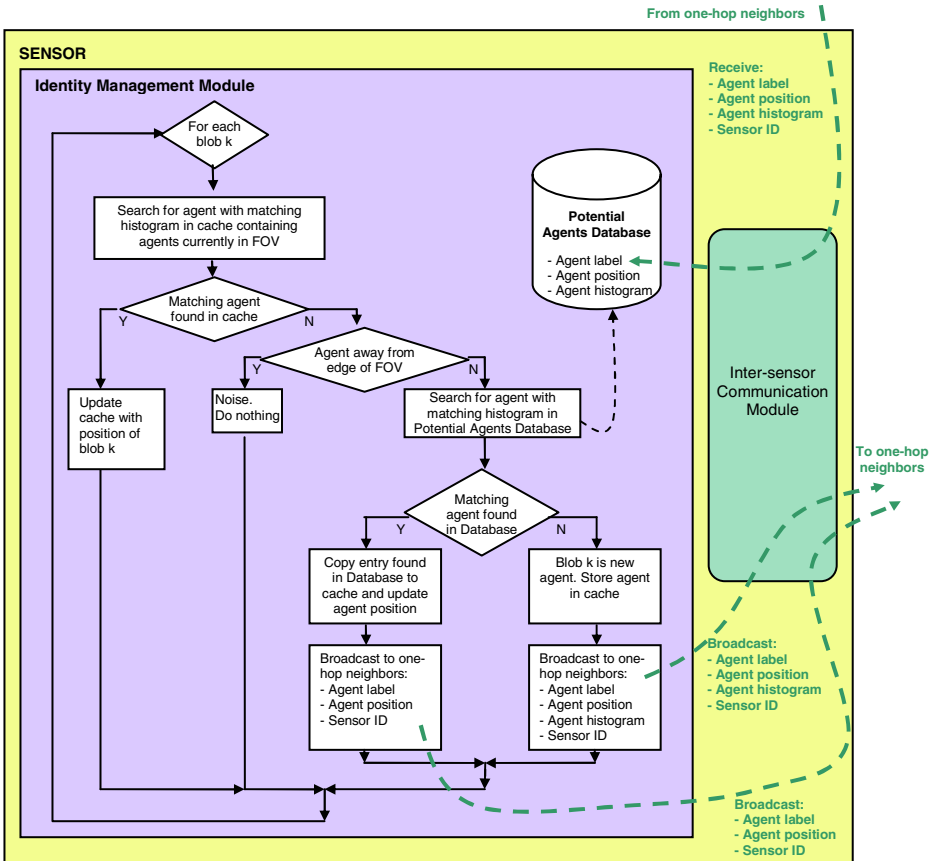


Fig. 3. Identity management algorithm

agent cannot be matched to any of those in the database, then it is declared as a new entry to the network, and the agent is stored in the cache and a copy is also stored in the database. The node then broadcasts a message to its one-hop neighbors, as described in the next section.

2.3 Inter-sensor Communication

Once a new agent is registered to the cache, the corresponding agent label is updated with the current timestamp, and this label along with the current position and the sensor ID is communicated to the one-hop neighbors. If the agent is also a new entry to the network, then the agent histogram is also transmitted.

This dissemination algorithm informs all neighboring nodes about the entry of an agent into the FOV of the sensor. Each sensor also listens to transmissions from its one-hop neighbors and keeps its “Potential Agents Database” up-to-date. Currently, the algorithm has been designed such that the sensor constantly listens for incoming packets.

A networked imaging system as devised here allows any end user at a sink node to make queries about some or all of the agents currently being tracked by the network. Specifically the user may wish to know the positions of the agents being tracked relative to a global coordinate system on the area encompassing the sensor network. Agent position information can be included in the packet sent to one-hop neighbors to allow for data aggregation through gateway nodes in a large sensor network. A query-based protocol can then collect the positions and identities of agents traveling in different FOVs for a centralized observer node.

To report agent positions, the coordinates of an agent relative to the FOV must be transformed into the global coordinate system. For an overhead camera system, the nodes could be localized using a method such as in Lee et al. [8], and the following transformation can be used to convert coordinates within a sensor's FOV to the global coordinates:

$$s_i = (\alpha R)^{-1} y_i + p . \quad (5)$$

In this equation, y_i are the coordinates of the agent in the FOV in which it is traveling, s_i are the coordinates in the global coordinate system, p is the vector extending from the origin of the global coordinate system to the origin of the sensor's FOV, α is the number of pixels in the FOV corresponding to one inch on the ground, and R is the rotation matrix with theta as the rotation in radians.

3 Results and Discussion

The networked tracking scheme described above operates under the assumption that the hue histograms of agents are invariant of the cameras and camera placements so that objects can be matched between different FOVs possibly observing different lighting conditions. This assumption was verified using a four-camera setup to compare the hue histogram produced by the different cameras. The effect of illumination on the hue histogram was investigated with a single camera and different illumination conditions. These experiments and their results are presented in the following sections.

3.1 Tracking Within the FOV

Fig. 4 illustrates an example of tracking two multi-colored agents traveling within the FOV of a camera. The displayed hue histograms are used to track the position of the agents as they appear in the subsequent frames captured by the camera.

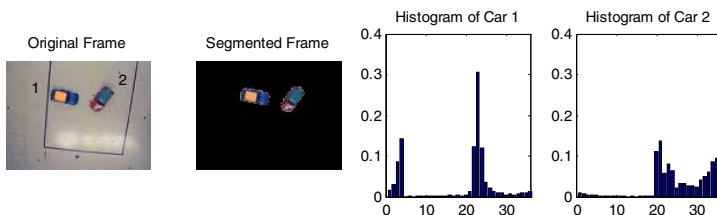


Fig. 4. Tracking multi-colored agents within the FOV of a camera using their hue histograms

3.2 Multiple FOV Operation

Three overhead cameras of the same model and one oblique camera of a different model were placed facing the floor of our lab, and three remote controlled cars were driven through their FOVs. The results of the experiment are as in Fig. 5.

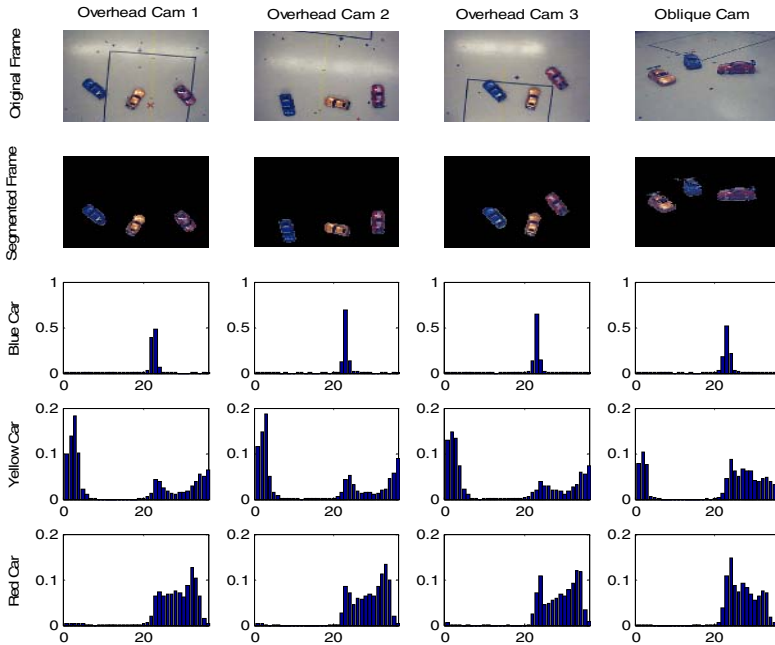


Fig. 5. Comparison of hue histograms produced by four different cameras for three agents. Each bin on the x-axis corresponds to ten degrees of hue.

The first, second, and third cars were blue, red, and yellow, respectively, but each car had details in other colors. It was observed that within a single FOV, both the overhead cameras as well as the oblique camera tracked the objects reliably. It is worth noting that the oblique camera tracked the three cars successfully regardless of perspective effects. However, the oblique camera could only track agents successfully under the assumption that the agents did not occlude each other. Occlusions could be resolved by the collaboration of two or more cameras, which has not been addressed in this work. Frames were captured from the four cameras and hue histograms were produced for each of the segmented blobs as presented in Fig. 5. The histograms obtained through this experiment demonstrate that each car's hue histogram retained its form between the cameras, despite the fact that the fourth camera had an oblique view and was of a different model.

3.3 Variations in Illumination Condition

The variation of the hue histogram within a single FOV with different lighting conditions was investigated for the same three agents presented in Sec. 3.2, and it was

observed that for all the three agents the histogram remained almost the same with varying illumination level. Fig. 6 presents the hue and intensity histograms for the yellow agent under three different illumination settings.

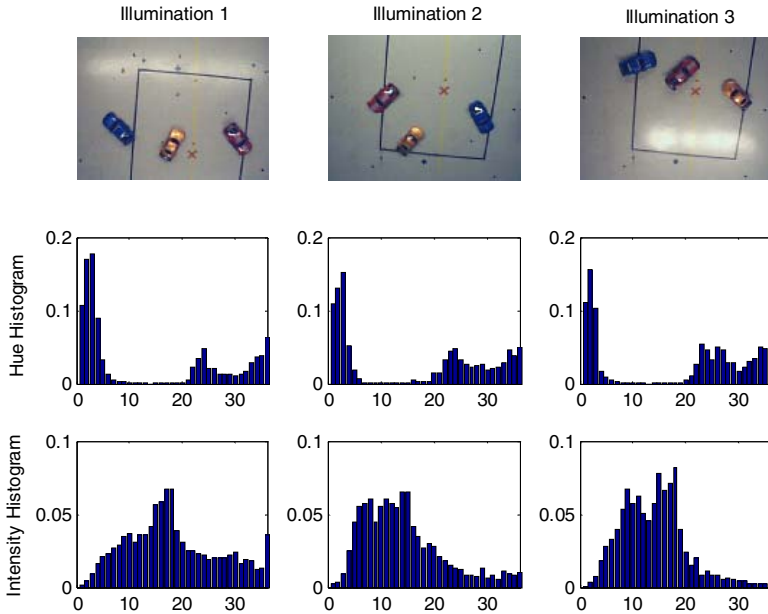


Fig. 6. Comparison of hue and intensity histograms for the yellow agent produced in three different illumination conditions. Each bin on the x-axis corresponds to ten degrees of hue.

The intensity histograms change depending on the illumination level. Therefore, by only considering the hue, the algorithm demonstrates robustness against changes in illumination. We did not have precise control over the illumination of the room and could not investigate the effect of larger changes in illumination. However, it was observed that the algorithm fails under very low illumination, such as when all the lights in the room are turned off and there is only a small amount of light entering through the door. It also fails under very heterogeneous illumination, that is, when part of the FOV has extremely low or extremely bright lighting. This result is expected since it is known from [9] that the hue histogram is only a definitive measure when the saturation is relatively high and the intensity is not extremely low or high. When the saturation is very low, the color becomes very pale and it is difficult to detect the hue. When the intensity is extremely low or high, the image is so dark or bright, respectively, that the hue is not apparent. In these cases, a color is better represented by its grayscale intensity rather than its hue. A possible solution to this problem is proposed in Sec. 4.

3.4 Collision Handling

COBMAT differentiates between agents after a collision with the same method as when no collisions occur, namely, by matching the hue histograms. Therefore, the elegance of our solution is that we use one technique to handle both tracking without collisions and

tracking with collisions. This ability to handle collisions is the primary advantage of attribute-based tracking versus path history-based tracking.

The current implementation of COBMAT cannot track multi-colored agents during a collision but can recover the agent identities right after the collision. If the collision occurs at the edge of the FOV, the sensor assumes that the merged blob is a new agent and broadcasts it to neighboring nodes, causing a false alarm. That condition needs special handling via predicting a collision when two agents approach each other in an edge zone, and is the subject of further investigation. However, by constraining the agents to single-colored objects, we were able to reliably track the agents even during collisions. Figure 7(a) shows the situation that occurs when agents collide: The blobs merge. Figure 7(b) illustrates how a tracking scheme would resolve the collision by taking the following steps to localize the agents within the merged blob.

- 1) Create a histogram for the merged blob.
- 2) Determine the number of peaks. Since the agents are single-colored, the number of peaks equals the number of agents in the merged blob (see Fig. 8(a)). Record the hue value of these peaks.
- 3) Divide the merged blob into sub-blocks (see Fig. 8(b)).
- 4) Create one histogram for each sub-block.
- 5) For each hue value recorded in step 2, determine which sub-block has a histogram with the greatest amplitude at this hue. The center of this sub-block is reported as the center of the agent with that hue. This is a reasonable estimate since the sub-block with the largest amplitude at a certain hue has the largest area that is of that hue.

An experimental test result for the case of single-color agents is depicted in Fig. 9.

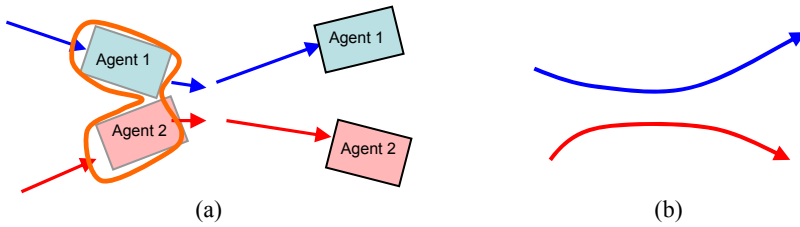


Fig. 7. (a) Merging of blobs in a collision (b) Resolved paths during a collision

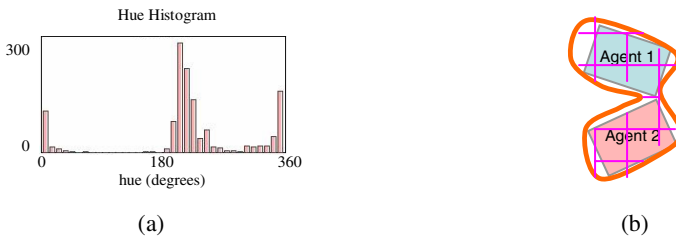


Fig. 8. (a) Two peaks in histogram indicate two single-color agents merged (peaks at 0 and 360 degrees both correspond to the red agent). (b) Scheme for localizing agents within merged blob.

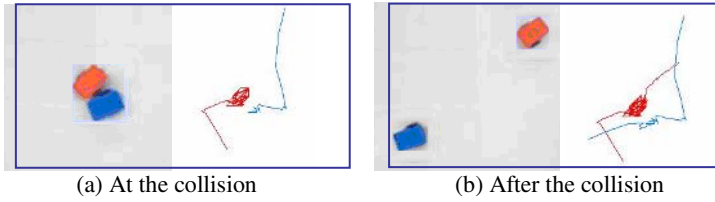


Fig. 9. Experimental result of the collision handling scheme for single-color agents, showing an illustration of the path plotted with the coordinates obtained

3.5 Computational Complexity

Since this algorithm is targeted for wireless sensor networks with low-power and low-cost sensor nodes, low computational complexity is one of the main objectives.

This algorithm achieves this goal by employing simple and elegant techniques designed with the specific needs of wireless sensor networks in mind. The computational complexity C for each sensor to analyze one frame is approximately

$$C = \alpha \cdot numAgents \cdot (\beta \cdot agentSize + \gamma \cdot numBins) + \kappa \cdot imageSize, \quad (6)$$

where $numAgents$ is the number of agents in the FOV, $agentSize$ is the size of each agent (in pixels), $numBins$ is the number of bins used in the histograms, $imageSize$ is the size of the FOV (in pixels), and α , β , γ , and κ are constants. κ accounts for the processing required to extract the blobs and γ represents the histogram matching computations. This algorithm is remarkably efficient because it is only linearly proportional to its parameters. Due to the broadcasting nature of transmissions, the network's operational complexity is also linear in the number of nodes. This makes the algorithm scalable, and hence feasible in large-scale wireless sensor networks.

4 Conclusions

A lightweight technique based on color histograms has been proposed for tracking multiple agents in a distributed image sensor network. The algorithm employs simple image processing and limited data communication between the nodes. The algorithm operates by exploiting the invariance of the hue histogram in different camera and different illumination settings to differentiate and match agents traveling within the network. It has been demonstrated that the hue histogram can be used reliably when the illumination within the FOV does not show drastic changes on the background. To adapt the current scheme to backgrounds with large illumination gradients, e.g. the floor of a forest, the histogram matching algorithm must be made adaptive to consider larger shifts of the hue histogram depending on the variation in the corresponding intensity histograms. It is also known that the hue histogram is not a reliable metric under very low illumination. To alleviate this effect, it may be possible to use hue and intensity histograms interchangeably depending on the saturation level of the image, stemming from the ideas presented in [9] and discussed in Sec. 3.3. However, the additional use of the intensity histogram will double the data overhead incurred.

References

1. Marcenaro, L., Ferrari, M., Marchesotti, L., Regazzoni, C.S., Multiple object tracking under heavy occlusions by using kalman filters based on shape matching. *Int. Conf. on Image Processing*, Rochester, NY (2002) 341–344
2. Nguyen, N., Bui, H.H., Venkatesh, S., West, G., Multiple camera coordination in a surveillance system. *ACTA Automatica Sinica*, Vol. 29 (2003) 408–422
3. Chang, T.H., Gong, S., Tracking Multiple People with a Multi-Camera System. *IEEE Workshop on Multi-Object Tracking* (2001) 19–26
4. Chang, C.K., Huang, J., Video surveillance for hazardous conditions using sensor networks. *IEEE Int. Conf. on Networking, Sensing & Control*, Taipei, Taiwan (2004) 1008–1013
5. Liu, F., Liu, Q., Lu, H., Robust Color-Based Tracking. *3rd Int. Conf. on Image and Graphics* (2004) 132–135
6. Limin, X., Object tracking using color-based kalman particle filters. *IEEE Int. Conf. on Signal Processing*, Beijing, China (2004) 679–682
7. Pérez, P., Hue, C., Vermaak, J., and Gangnet, M., Color-Based Probabilistic Tracking. *7th European Conf. on Computer Vision-Part I* (2002) 661–674
8. Lee, H., Aghajan, H., Vision-Enabled Node Localization in Wireless Sensor Networks. *COGNITIVE systems with Interactive Sensors*, Paris, France (2006)
9. Sural, S., Qian, G., Pramanik, S., A histogram with perceptually smooth color transition for image retrieval. *4th Int. Conf. on Computer Vision, Pattern Recognition and Image Processing*, Durham (2002)

A Fast Motion Vector Search Algorithm for Variable Blocks

Yung-Lyul Lee¹, Yung-Ki Lee¹, and HyunWook Park²

¹ Sejong University, Department of Internet Engineering, DMS Lab.,
98 Kunja-dong, Kwangjin-gu, Seoul, Korea
yllee@sejong.ac.kr

² Department of Electrical Engineering
KAIST, 373-1 Guseong-dong, Yuseong-gu, Daejeon 305-701, Korea

Abstract. A fast motion estimation (ME) algorithm is proposed to search motion vectors for variable blocks. The proposed method is based on the successive elimination algorithm (SEA) using sum norms to find the best estimate of the motion vectors and to implement efficient calculations for variable blocks. The proposed ME algorithm is applied to the Joint Video Team (JVT) encoder that performs a variable-block ME. In terms of computational complexity, the proposed ME algorithm with limited search range searches motion vectors at about 6.3 times as fast as the spiral full search and 5.5 times as fast as the fast full search using the hierarchical sum of absolute difference (SAD), while the PSNR (peak signal-to-noise ratio) of the reconstructed image is slightly degraded with only 0.1~0.4 dB.

1 Introduction

Most video compression standards, including ITU-T H.263 [1], MPEG-4 [2], and the JVT codec [3] which is a joint standard of ITU-T Recommendation H.264 and ISO/IEC MPEG-4 Part 10, use a block-based motion estimation/compensation. Motion estimation (ME) plays an important role in saving bit-rates by reducing temporal redundancy in moving picture coding, but it requires intensive computations in the encoder. In particular, the JVT codec adopts the variable block-based motion estimation/compensation for each 16×16 macroblock (MB), whose motion vectors have quarter pixel resolution. It improves the peak signal to noise ratio (PSNR) and subjective quality, but it requires heavy computations for motion estimation.

Several ME algorithms have been proposed to save computation time. The well-known approaches to speed up motion estimation are the three-step search [4], the 2-D logarithmic search [5], the one-at-a-time search (OTS) [6], and the new diamond search [7]. These methods estimate integer motion vector with approximately 3%~5% of the computational complexity compared to the full search by sampling of the search positions for ME. However, these approaches cannot provide as high PSNR as the full search. The successive elimination algorithm (SEA) [8] shows the same PSNR as the full search with 13% computational complexity of the full search. The enhanced SEA (ESEA) [9] further reduces computational complexity, while maintain-

ing the PSNR. The spiral full search [10] in the JM (Joint Model) of JVT performs a full search in a way that an initial motion vector is predicted by median value of motion vectors of the adjacent blocks and the motion vector search is performed in the spiral sequence from the median-predicted motion vector to ± 16 motion vector. The fast full search [10] in the JM performs SAD calculations of sixteen 4×4 blocks in an MB over ± 16 search ranges and the SAD values of 4×4 blocks are used hierarchically to estimate the motion vectors of larger blocks.

In order to apply the concept of SEA to the JVT codec, a new fast motion vector search algorithm is proposed for integer-pixel unit. We also use the rate-distortion optimization (RDO) [11] that was adopted in JVT as an informative module. The RDO is applied to select the optimum variable block size, with which we have both the minimum mean-square error and the minimum bit allocations for each 16×16 MB.

This paper consists of the following three sections. In Section 2, a detailed proposed method is described. Experimental results are shown to evaluate the proposed method in Section 3. Finally, conclusions are given in Section 4.

2 Motion Estimation Algorithm in Integer-Pixel and Quarter-Pixel Units

2.1 Motion Estimation Algorithm in Integer-Pixel Unit

The JVT codec uses the motion estimation of the variable blocks such as 16×16 , 16×8 , 8×16 , and 8×8 blocks for each 16×16 MB and 8×4 , 4×8 and 4×4 blocks for each 8×8 block. In the hierarchical motion vector search algorithm of the variable blocks, the best motion vector is found in consideration of the number of bits to represent motion vectors and the SAD of the motion-compensated MB. In this paper, sum norms are calculated in advance on the reference frame for fast motion search.

Assume that the current frame is denoted by $f(i,j)$ with spatial coordinates (i,j) , and the reference frame is $r(i,j)$. Motion estimation of $S \times T$ blocks, such as 16×16 , 16×8 , 8×16 , 8×8 , 8×4 , 4×8 , and 4×4 , is performed by using the inequality equation [8, 12] as follows:

$$\left| \sum_{i=1}^S \sum_{j=1}^T |r(i-x, j-y)| - \sum_{i=1}^S \sum_{j=1}^T |f(i, j)| \right| \leq \sum_{i=1}^S \sum_{j=1}^T |r(i-x, j-y) - f(i, j)| \quad (1)$$

where (x, y) represents a motion vector. For simplicity, eq. (1) is described as follows:

$$\left| R_{S \times T}(x, y) - F_{S \times T} \right| \leq SAD_{S \times T}(x, y) \quad (2)$$

$$R_{S \times T}(x, y) = \sum_{i=1}^S \sum_{j=1}^T |r(i-x, j-y)|$$

where $F_{S \times T} = \sum_{i=1}^S \sum_{j=1}^T |f(i, j)|$

$$SAD_{S \times T}(x, y) = \sum_{i=1}^S \sum_{j=1}^T |r(i-x, j-y) - f(i, j)|$$

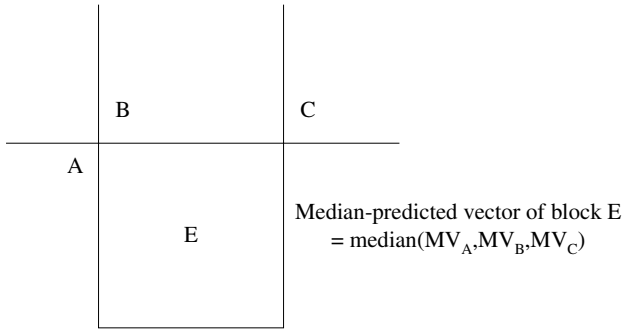


Fig. 1. Neighboring blocks for the motion vector prediction of current block. E is the current block and A, B, and C are the neighboring blocks. MV_A , MV_B , and MV_C are motion vectors of the neighboring blocks A, B, and C, respectively.

In eq. (2), $R_{S \times T}(x, y)$ is the sum norm of $S \times T$ block of the reference frame with the motion vector (x, y) , $F_{S \times T}$ is the sum norm of $S \times T$ block of the current frame and $SAD_{S \times T}(x, y)$ is the sum of absolute difference (SAD) between the current $S \times T$ block and the reference $S \times T$ block with the motion vector (x, y) . If the motion vector (x, y) is a better estimation than the other motion vector (m, n) in terms of SAD, then the inequality is given as follows:

$$SAD_{S \times T}(x, y) \leq SAD_{S \times T}(m, n) \tag{3}$$

Eqs. (2) and (3) can be combined as follows:

$$|R_{S \times T}(x, y) - F_{S \times T}| \leq SAD_{S \times T}(m, n) \tag{4}$$

Eq. (4) can be rearranged without absolute terms as follows:

$$F_{S \times T} - SAD_{S \times T}(m, n) \leq R_{S \times T}(x, y) \leq F_{S \times T} + SAD_{S \times T}(m, n) \tag{5}$$

If eq. (5) is satisfied on the motion vector (x, y) , $SAD_{S \times T}(x, y)$ is calculated and $SAD_{S \times T}(m, n)$ is replaced with $SAD_{S \times T}(x, y)$. To obtain the best estimation of the current $S \times T$ block, computation of SAD is performed only for the motion vector (x, y) satisfying eq. (5). The number of motion vectors satisfying eq. (5) is obviously less than the total search range. Therefore, the proposed algorithm can reduce the computational complexity without degradation of the PSNR. The efficiency of the algorithm depends on the fast calculation of the sum norms for each block and the initial motion prediction, $SAD_{S \times T}(m, n)$. Fig. 1 illustrates the relative location of the current block E and the neighboring blocks of A, B, and C, whatever the block size is. The median-predicted motion vector PMV is used as the initial motion vector of the current block, which is predicted from the upper block B, the upper right block C, and the left block A of the current block E as follows:

$$PMV = \text{median}(MV_A, MV_B, MV_C),$$

where MV_A , MV_B , and MV_C are motion vectors of A, B, and C, respectively in Fig. 1.

Since JVT adopts variable blocks for motion estimation/compensation, there are various numbers of motion vectors for each MB according to the block size. As an example, if all 4×4 blocks in an MB are selected, 16 motion vectors should be encoded and transmitted to receiver. Including the motion vector cost as suggested in the JVT encoder, eq. (3) is modified as follows:

$$\begin{aligned} SAD_{S \times T}(x, y) + \lambda \times MVbits(PMV - (x, y)) \\ \leq SAD_{S \times T}(m, n) + \lambda \times MVbits(PMV - (m, n)) \end{aligned} \tag{6}$$

where the Lagrangian multiplier λ is set to $\sqrt{0.85 \times 2^{(QP-12)/3}}$ as defined in the JVT encoder [3, 10, 11], $MVbits(PMV - (x, y))$ is the bit amount to represent the difference between the motion vector (x, y) and the predicted-motion vector PMV by Exp-Golomb code, and QP is the quantization parameter of the JVT codec. Then, eq. (6) is rearranged to be used for the proposed fast motion estimation as follows:

$$\begin{aligned} F_{S \times T} - (SAD_{S \times T}(m, n) + \lambda \times (MVbits(PMV - (m, n)) - MVbits(PMV - (x, y)))) \\ \leq R_{S \times T}(x, y) \\ \leq F_{S \times T} + (SAD_{S \times T}(m, n) + \lambda \times (MVbits(PMV - (m, n)) - MVbits(PMV - (x, y)))) \end{aligned} \tag{7}$$

Eq. (7) is the proposed sum norm inequality which considers both SAD and the motion vector cost. Weighting of SAD and the motion vector cost can be adjusted by the Lagrangian multiplier λ , which is the same value as JVT encoder in this paper.

2.2 Efficient Calculation of the Hierarchical Sum Norms of Variable Blocks

The sum norms of the variable blocks are computed hierarchically. First of all, the sum norms of the smallest blocks of 4×4 among variable blocks are first calculated in the reference frame. In order to reduce the computation amount, an efficient procedure to compute the sum norms of 4×4 blocks is described. The first row strip, $C(0, j)$, that is the sum of four elements of j -th row is computed for reference frame as follows:

$$C(0, j) = \sum_{k=0}^3 r(k, j) \text{ for } 0 \leq j < W \tag{8}$$

Then, next row strip, $C(i, j)$ for $i \geq 1$, is computed by using $C(i-1, j)$ as follows:

$$C(i, j) = C(i-1, j) - r(i-1, j) + r(i+3, j) \text{ for } 1 \leq i < H-3, 0 \leq j < W \tag{9}$$

From $C(i, j)$, $R_{4 \times 4}(x, y)$ ($x=0, 1, 2, \dots, H-4$, $y=0, 1, 2, \dots, W-4$) is derived in the same way as the sum of four column elements as follow:

$$\begin{aligned} R_{4 \times 4}(i, 0) &= \sum_{k=0}^3 C(i, k) \text{ for } 0 \leq i < H-3 \\ R_{4 \times 4}(i, j) &= R_{4 \times 4}(i, j-1) - C(i, j-1) + C(i, j+3) \text{ for } 0 \\ &\leq i < H-3, 1 \leq j < W-3 \end{aligned} \tag{10}$$

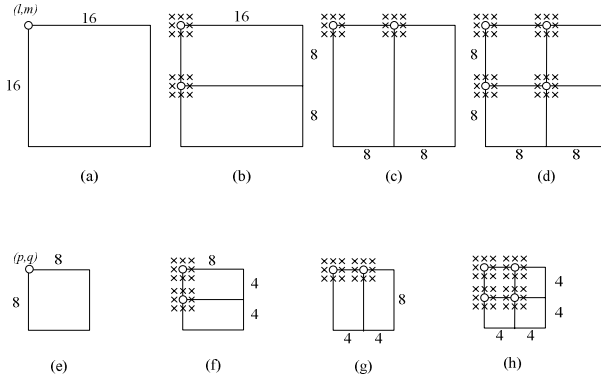


Fig. 2. Limited search ranges for each variable block matching: (a) matching position of the 16×16 macroblock in the reference frame, (b) search positions in the reference frame of the two 16×8 blocks, (c) search positions in the reference frame of the two 8×16 blocks, (d) search positions in the reference frame of four 8×8 blocks, (e) matching position of the 8×8 block in the reference frame, (f) search positions in the reference frame of two 8×4 blocks, (g) search positions in the reference frame of two 4×8 blocks, (h) search positions in the reference frame of four 4×4 blocks

From the above sum norms of 4×4 blocks, the sum norms of 4×8 and 8×4 blocks are calculated as follows:

$$R_{4 \times 8}(i, j) = R_{4 \times 4}(i, j) + R_{4 \times 4}(i, j + 4) \text{ for } 0 \leq i < H - 3, \quad 0 \leq j < W - 7 \tag{11}$$

$$R_{8 \times 4}(i, j) = R_{4 \times 4}(i, j) + R_{4 \times 4}(i + 4, j) \text{ for } 0 \leq i < H - 7, \quad 0 \leq j < W - 3 \tag{12}$$

Also, $R_{8 \times 8}(i, j)$, $R_{8 \times 16}(i, j)$, $R_{16 \times 8}(i, j)$, and $R_{16 \times 16}(i, j)$ can be derived in the same way as $R_{4 \times 8}(i, j)$ and $R_{8 \times 4}(i, j)$. The proposed hierarchical sum norm calculation is based on 4×4 sum norm and expanded to the sum norms calculation of larger blocks using 4×4 sum norm in eqs. (11) and (12). The main contributions on the proposed method are that the proposed sum norm inequality of eq. (7) considers both SAD and the motion vector cost and the proposed hierarchical sum norms calculations of eqs. (8)~(11) enable to find the motion vector very fast.

2.3 Motion Estimation Procedure for Variable Blocks in Integer-Pixel Unit

The hierarchical calculation of sum norms can reduce the computation time. The sequence of variable block matching is executed from the 16×16 block to the 4×4 blocks to find the best estimates of the motion vectors and block sizes. After obtaining the seven sum-norm sets of $R_{4 \times 4}(i, j)$, $R_{4 \times 8}(i, j)$, $R_{8 \times 4}(i, j)$, $R_{8 \times 8}(i, j)$, $R_{8 \times 16}(i, j)$, $R_{16 \times 8}(i, j)$, and $R_{16 \times 16}(i, j)$ in the reference frame, the motion estimation of variable blocks is performed by eq. (7) for 16×16 MBs. Once we find the motion vector (l,m) that has a minimum SAD for 16×16 MB as shown in Fig. 2(a), motion estimation of 16×8, 8×16, 8×8 blocks are performed for the motion vector (l,m) and its eight-neighbor search ranges as shown in Fig. 2(b)-(d), in which the eight-neighbor search

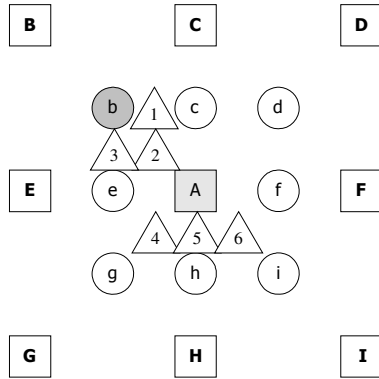


Fig. 3. A fast quarter pixel search method: the upper-case letters are integer-pixel unit, the lower case letters are half-pixel unit, and the numbers in triangle are quarter-pixel unit

ranges mean ± 1 integer search range around the motion vector (l,m) for estimating 16×8 , 8×16 , 8×8 block motion vector. This limited search range can reduce the computation amount, whereas image quality is slightly degraded. After we find an optimum motion vector for each 8×8 block, the 8×8 block is divided into two 8×4 blocks, two 4×8 blocks and four 4×4 blocks. If the optimum motion vector of the 8×8 block is (p,q), the motion estimation of 4×8 , 8×4 , and 4×4 blocks is performed for the motion vector (p,q) and its eight neighbor motion vectors. In order to retain the image quality, the search range can be ± 16 for all variable blocks, which requires a longer computation time compared to the limited search range.

The proposed algorithm adopts the early termination to reduce the computations of $SAD_{S \times T}(x, y)$. $SAD_{S \times T}(x, y)$ should be calculated only if eq. (7) is satisfied. When $SAD_{S \times T}(x, y)$ is calculated, the intermediate result of $SAD_{S \times T}(x, y)$ between the current block and the reference block after 75% calculations of $S \times T$ block is compared with $SAD_{S \times T}(m, n)$. If the intermediate result is greater than $SAD_{S \times T}(m, n)$, the remaining calculation is not necessary. Therefore, it slightly contributes to the reduction of computation amounts.

2.4 Fast Quarter-Pixel Search

For fast quarter-pixel motion estimation, three points quarter-pixel search algorithm is developed. An example of quarter-pixel search is shown in Fig. 3, in which the upper-case letters are integer-pixel unit, the lower case letters are half-pixel unit, and the numbers in triangle are quarter-pixel unit. When an integer pixel is found as the matching block of a block that can be one of 16×16 , 16×8 , 8×16 , 8×8 , 8×4 , 4×8 , and 4×4 blocks according to the value of S and T, the eight-neighbor positions of b, c, d, e, f, g, h, and i in half pixel unit are investigated by full search and the candidate position that has the minimum MAD is selected as the best half pixel motion vector position. If the half-pixel motion vector position is b, only positions of 1, 2, and 3 in quarter-pixel unit are investigated for the quarter-pixel search. When the half-pixel

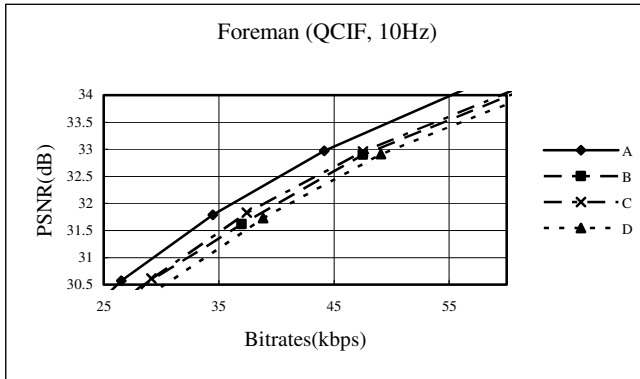
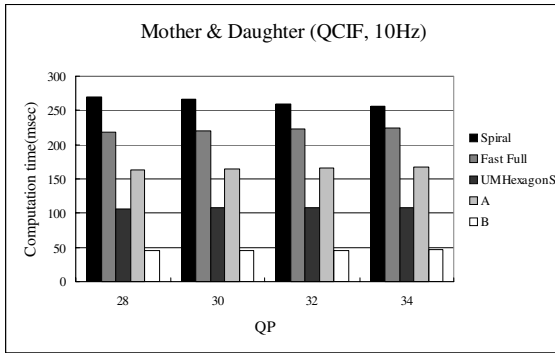


Fig. 4. Rate-distortion plots of the proposed methods “A” and “B”, the methods “C”, and “D” to compare the performance of eq. (5) and eq. (7)

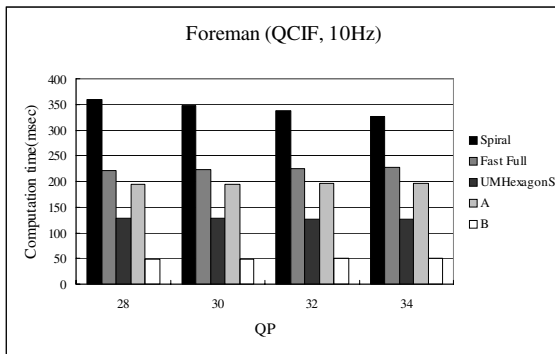
motion vector position is h , only positions of 4, 5, and 6 are investigated for the quarter-pixel search. That is, only three quarter pixel positions between the integer motion vector position and the half-pixel motion vector position are investigated as the candidate of quarter-pixel motion vector, because the possibility of the best quarter pixel candidate is very high within three points in quarter pixel unit between the best integer and best half pixel position.

3 Experimental Results

The proposed motion vector search algorithm is applied to the JVT encoder [3,10] with the Exp-Golomb code, variable-block ME/MC having 16×16 , 16×8 , 8×16 , 8×8 , 8×4 , 4×8 , and 4×4 blocks, ± 16 motion search range, quarter pixel interpolation, 4×4 DCT (Discrete Cosine Transform), and rate-distortion optimization. Several image sequences, each of which has 300 frames, were used for this experiment. Each sequence is compressed with the scheme of I,P,P,P,P... , i.e., only the first frame is INTRA frame, while the others are all INTER frames without the B frame. The proposed method is compared to the spiral full search, the fast full search and UMHexagon search (Unsymmetrical-cross Multi-Hexagon grid Search) [13], which are implemented in JM (Joint Model) source code [10], with respect to PSNR and computation time. In the comparison study, the “A” method is the proposed method using eq. (7) with ± 16 search ranges for all variable blocks that gives the same PSNR as the spiral full search, and the “B” method is the proposed method using eq. (7) with the limited search ranges described in Section II. The fast full search in the JM source code performs SAD calculations of 4×4 blocks for variable-block motion estimation, in which SAD values of 4×4 blocks are used hierarchically to estimate motion vectors of larger blocks. Also, the same rate distortion optimization technique, which was introduced in the references [10, 11], is used for fair comparison.



(a)

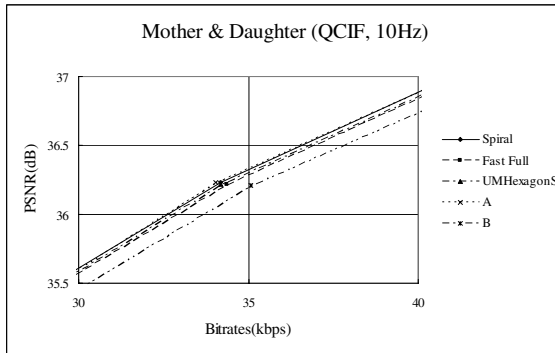


(b)

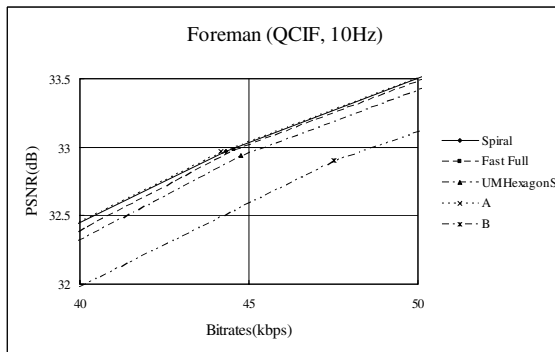
Fig. 5. The computation time of the spiral full search, the fast full search, UMHexagonS, the proposed method “A”, and “B”. Only the first frame is Intra-coded: (a) “Mother & Daughter” sequence with a frame rate of 10Hz and QCIF resolution, (b) “Foreman” sequence with a frame rate of 10Hz and QCIF resolution.

In order to show that eq. (7) has PSNR improvement in comparison to eq. (5), the motion estimation using eq. (7) is compared to that using eq. (5) in terms of PSNR as shown in Fig. 4. “C” and “D” are the methods using eq. (5) with ± 16 and the limited search ranges, respectively. The result of eq. (7) is approximately 0.4~0.5 dB better than that of eq. (5) in the Foreman QCIF sequence with a frame rate of 10 Hz.

All experiments were carried out on a Pentium IV, 1.7 GHz, using the JVT (Joint Video Team) codec [10] for various video sequences with QCIF and CIF size. The computation times of four different motion estimation methods are compared for two QCIF sequences in Figs. 5(a) and (b), where the horizontal axis represents the quantization parameters of the JVT codec. The computation time(msec) is the measured average time per a frame on the Pentium IV-1.7GHz for only motion estimation in JM73. The “A” method shows that the proposed method results in the same PSNR as the spiral full search if it is applied to ± 16 search ranges for all variable-blocks whereas computation amount increases in Figs. 5 and 6. UMHexagonS, which is



(a)



(b)

Fig. 6. Rate-distortion plots: (a) “Mother & Daughter” rate-distortion plot corresponding to Fig. 5(a), (b) “Foreman” rate-distortion plot corresponding to Fig. 5(b)

adopted as an informative part of JVT, is the motion estimation method using initial search point prediction and early termination. In the “Mother & Daughter” and “Foreman” sequences, Figs. 5(a) and (b), show that the proposed search method is approximately 6.3 times faster than the spiral full search method, 5.5 times faster than the fast full search, and 2.4 times faster than the UMHexagon search in the JM source code. Figs. 6(a) and (b) show the rate distortion curves of the “Mother & Daughter” and “Foreman” sequence of QCIF, in which a frame rate of 10 Hz was used in this experiment. In terms of rate-distortion curve, the PSNR of the proposed method is approximately 0.4dB degraded in comparison to that of the fast full search, the spiral full search and the UMHexagon search methods. If the image quality is more important than the computation time, we can increase the search range into 16×16 in the proposed method, which improves the PSNR. Figs. 7 and 8 show the computation times and the rate distortion curves of four motion estimation methods for the two CIF sequences, “Paris” and “Foreman”, with frame rates of 10 Hz. In the CIF sequences, the proposed method shows similar results to that of the QCIF sequences.

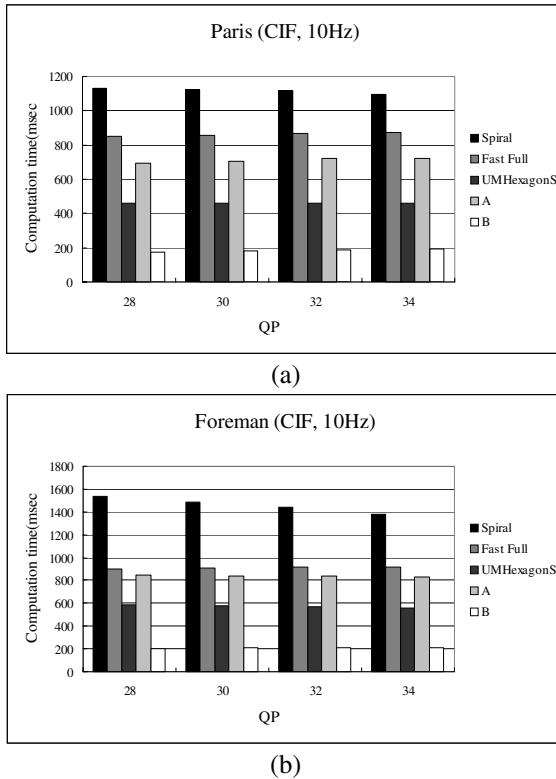
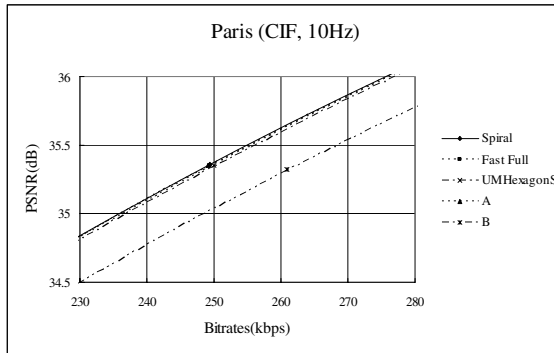


Fig. 7. The computation time of the spiral full search, the fast full search, UMHexagonS, and the proposed methods “A” and “B”. Only the first frame is Intra-coded: (a) “Paris” sequence with a frame rate of 10Hz and CIF resolution, (b) “Foreman” sequence with a frame rate of 10Hz and CIF resolution.

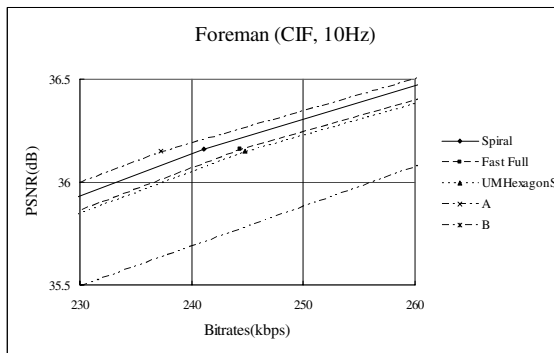
The “A” method always shows the best PSNR as shown in Figs. 6 and 8. The “B” method requires the smallest computation amount as shown in Figs. 5 and 7, whereas PSNR is degraded. Therefore, methods “A” or “B” can be applied alternately according to the importance of the image quality and the computation time.

4 Conclusions

The proposed motion vector search method utilizing hierarchical sum norm, which is developed through eqs (6)~(12) in this paper, can be easily applied to the JVT codec. The JVT codec, which is H.264 and MPEG-4 part-10 video coding standard, adopted variable blocks and quarter-pixel motion estimation/compensation in its codec. Therefore, the proposed method can be applied to the JVT codec to reduce the computational complexity of the search process with a very small loss in PSNR. The proposed one can also be applied to multiple reference frames without a serious loss of video quality. The hierarchical sum norm and the fast three points search for quarter-pixel motion estimation contribute to the reduction of computational complexity.



(a)



(b)

Fig. 8. Rate-distortion plots: (a) “Paris” rate-distortion plot corresponding to Fig. 7(a), (b) “Foreman” rate-distortion plot corresponding to Fig. 7(b)

References

- [1] ITU Telecom. Standardization Sector, “Video Codec Test Model Near-Term, Version 10 (TMN10) Draft 1,” *H.263 Ad Hoc Group*, April 1998.
- [2] “Information Technology - Coding of Audio-Visual Objects Part2: Visual Amendment 1: Visual Extensions,” *ISO/IEC JTC1/SC29/WG11 N3056*, Dec. 1999.
- [3] T. Wiegand, Final draft international standard for joint video specification H.264, in *JVT of ISO/IEC MPEG and ITU-T VCEG*, JVT-G050, Mar. 2003.
- [4] T. Koga, K. Iinuma, A. Hirano, Y. Iijima, and T. Ishiguro, “Motion compensated inter-frame coding for video conferencing,” in *Proc. Nat. Telecommunications Conf.*, New Orleans, LA, pp. G.5.3.1-G.5.3.5., Nov. 1981.
- [5] J. R. Jain and A. K. Jain, “Displacement measurement and its application in interframe image coding,” *IEEE Trans. Commun.*, vol. 29, pp.1799-1808, Dec. 1981.
- [6] R. Srinivansan and K. Rao, “Predictive coding based on efficient motion estimation,” *IEEE Trans. Commun.*, vol. COM-33, pp. 1011-1015, Sept. 1985.
- [7] S. Zhu and K-K. Ma, “A New Diamond Search Algorithm for Fast Block-Matching Motion Estimation,” *IEEE Trans. Image Processing*, vol. 9, no. 2, pp.287-290, Feb. 2000.

- [8] W. Li and E. Salari, "Successive Elimination Algorithm for Motion Estimation," *IEEE Trans. Image Processing*, vol. 9, no. 1, pp.105-107, Jan. 1995.
- [9] M. Brunig and W. Niehsen, "Fast Full-Search Block Matching," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 11, no. 2, pp. 241-247, Feb. 2001.
- [10] http://bs.hhi.de/~suehring/tml/download/old_jm/jm73.zip
- [11] G. Sullivan and T. Wiegand, "Rate-Distortion Optimization for Video Compression," *IEEE Signal Processing Magazine*, pp. 74-90, Nov. 1998.
- [12] D. M. Young and R. T. Gregory, *A Survey of Numerical Mathematics*, New York: Dover, vol. 2, pp. 759-762, 1988.
- [13] Z. Chen, P. Zhou, and Y. He, "Fast Motion Estimation for JVT," *JVT of ISO/IEC MPEG & ITU-T VCEG*, JVT-G016, Mar. 2003.

Constrained Region-Growing and Edge Enhancement Towards Automated Semantic Video Object Segmentation

L. Gao¹, J. Jiang², and S.Y. Yang¹

¹Institute of Acoustics, Chinese Academy of Sciences, China

²School of Informatics, University of Bradford, UK

Future_Gao@hotmail.com, J.Jiang1@brad.ac.uk,

S.Y.Yang@hotmail.com

Abstract. Most existing object segmentation algorithms suffer from a so-called under-segmentation problem, where parts of the segmented object are missing and holes often occur inside the object region. This problem becomes even more serious when the object pixels have similar intensity values as that of backgrounds. To resolve the problem, we propose a constrained region-growing and contrast enhancement to recover those missing parts and fill in the holes inside the segmented objects. Our proposed scheme consists of three elements: (i) a simple linear transform for contrast enhancement to enable stronger edge detection; (ii) an 8-connected linking regional filter for noise removal; and (iii) a constrained region-growing for elimination of those internal holes. Our experiments show that the proposed scheme is effective towards revolving the under-segmentation problem, in which a representative existing algorithm with edge-map based segmentation technique is used as our benchmark.

1 Introduction

Video object segmentation consistent with human visual perception has long been identified as a difficult problem since it requires characterization of semantics to define the objects of interest. Unique definition of such semantics is not possible as the semantics are often context dependent and thus low-level segmentation has been focusing on segmentation of regions rather than objects. As video processing and coding is moving towards content-based approaches, object segmentation becomes an important research topic. The content-based video compression standard, MPEG-4, stands as a most representative scenario for such content-based approaches. As a result, research on this subject has been very active and many algorithms have been reported in the literature. Existing video object segmentation can be roughly classified into two categories according to their primary segmentation criteria. One category is represented by those regional segmentation techniques [1-5], where spatial homogeneity is primarily used as the criteria to develop rules towards the segmentation design. As these techniques are rooted among low-level image processing and essentially data-driven, precise boundaries of the segmented regions can often be obtained. However the computation incurred is normally high since iterative operations are often required. Examples of such techniques include watershed, snake modeling, and region-growing [11] etc. The other category of segmentation can be characterized by detection of changes [6-15], where motion information is utilized to segment those

moving objects together with other spatio-temporal information. In this category, the objects segmented are close to semantic video objects and thus provide promising platforms for further research and development.

Specifically, existing research on video object segmentation is built upon change detection assisted by other sideline information including spatial segmentation, edge detection and background registration etc. [6-15]. In [6], Kim et. al. described a spatio-temporal approach for automatic segmentation of video objects, where hypothesis test based on estimated variances within a window is proposed to exploit the temporal information, and spatial segmentation is included to assist with detection of object boundaries. The final decision on foreground and background objects is made in combining the spatially segmented object mask with the temporally segmented object mask, in which a two-stage process is designed to consider both the change detection and the historical attributes. In [7], another similar approach was described towards a robust or noise-insensitive video object segmentation, which follows the idea of combining spatial edge information with motion-based edge detection. Based on these algorithms, we carried out a series of empirical studies and testing. Our experiments reveal that, while the algorithms perform well generally, there exist an under-segmentation problem, where parts of the object region are missing or there exist holes inside the object region. This is a serious problem especially when the background has similar intensity values to those inside objects or at the boundary of the object region. To rectify this problem, we propose a new scheme of automatic semantic object segmentation, where elements of edge enhancement and constrained region growing are proposed, combing the strength of change detection with the strength of spatial segmentation (region-grow). We also illustrate via extensive experiments how our proposed algorithm could achieve this objective in comparison with the existing VO segmentation algorithm reported in [7]. The rest of the paper is organized into two sections. One section is dedicated to our proposed algorithm design, and the other is dedicated to experiments and presentation of their results. Finally some concluding remarks will also be made in the same section.

2 The Proposed Algorithm Design

2.1 Edge Enhancement and Linear Filtering

In practice, when the grey level difference between the object and the background is small, part of the object at its boundary will have similar intensity values to that of background. In this circumstance, edge detection could fail to detect all the edges of the object, and thus some parts inside the objects become missing. To reduce such a negative effect upon object segmentation, we propose a simple linear transformation as part of the pre-processing to enhance the contrast of the luminance component of the video frame before edge detection is performed. Although there exist many contrast enhancement algorithms that may provide better performances, our primary aim here is not only improving the segmentation accuracy, but also maintaining the necessary simplicity for real-time applications. Considering the fact that increase of

contrast will inevitably introduce additional noise, we also designed a simple 2D filter to remove the noise. By combining these two elements together, we achieve the objective that edges are enhanced to enable edge detection to extract the boundaries of moving objects and hence ready for semantic object segmentation.

Given an input video frame, $I_n(x, y)$, assuming that their intensity values are limited to the range of $[a, b]$, its transformed video frame can be generated as:

$$g_n(x, y) = a' + \frac{b' - a'}{b - a} \times (I_n(x, y) - a) \tag{1}$$

Where $a' = 0, b' = 255$ and $a = 70, b = 180$.

Following the contrast enhancement, we then apply Canny edge detector [7] to extract edges from the video frames to characterize the semantic objects to be segmented. This is essentially a gradient operation performed on the Gaussian convoluted image. Given the n th video frame I_n , the Canny edge detecting operation can be represented as:

$$\Phi(I_n) = \theta(\nabla G * I_n) \tag{2}$$

where $G * I_n$ stands for the Gaussian convoluted image, ∇ for the gradient operation, and θ for the application of non-maximum suppression and the thresholding operation with hysteresis to detect and link the edges.

Following the spirit of the work reported in [7], we extract three edge maps: (i) the difference edge map $DE_n = \Phi(I_{n-1} - I_n)$, (ii) the current edge map $E_n = \Phi(I_n)$, and (iii) background edge map E_b , which contains background edges to be defined by manual process or by counting the number of edge occurrences for each pixel through the first several frames [7]. Our implementation adopted the latter option.

From these three edge maps, a currently moving edge map, ME_n^{change} , representing the detected changes can be produced as follows:

$$ME_n^{change} = \left\{ e \in E_n \mid \min_{x \in DE_n} \|e - x\| \leq T_{change} \right\} \tag{3}$$

where e stands for edge pixels inside the moving edge map, and T_{change} for a threshold empirically determined as 1 in [7]. Essentially, (3) describes an operation in selecting all edge pixels within a small distance of DE_n .

Further, a temporarily still moving edge map ME_n^{still} can also be produced by considering the previous frame's moving edges. This edge map is used to characterize those regions that belong to the moving object but temporally no change is incurred between two adjacent frames. Such an operation can be described as given below:

$$ME_n^{still} = \left\{ e \in E_n \mid e \notin E_b, \min_{x \in ME_{n-1}} \|e - x\| \leq T_{still} \right\} \tag{4}$$

where T_{still} is another threshold, which is also empirically determined as 1 in [7]. As indicated in (3), temporarily still moving edge map contains edge pixels that they are part of current edge map but not part of background edge map, and they also satisfy the condition: $\min_{x \in ME_{n-1}} \|e - x\| \leq T_{still}$.

After the identification of those moving edges by (3) and (4), the remaining operation for extracting video objects is to combine the two edge maps into a final moving edge map: $ME_n = ME_n^{change} \cup ME_n^{still}$, and then select the object pixels via a logic AND operation of those pixels between the first and the last edge pixel in both rows and columns [7].

As the contrast enhancement introduced by (1) often produce noise, the edge maps produced could be affected. As a result, to remove the additional noise introduced by the linear transformation, we adopt the method of 8-connected linking region sign to design a filter and apply the filter to both the motion edge map ME_n and the extracted object sequences.

Given an edge pixel at (x, y) whose value is 1 in the binary map, we examine its 3×3 neighborhood to produce a set of all connected points $N = \{A_1, A_2 \dots A_k\}$. If $k \leq T$, a threshold for noise removal, the set of points N will be regarded as noise and thus deleted.

2.2 Constrained Region Growing

Most existing segmentation algorithms often include post-processing to improve the final segmented objects [7-15]. The change detection based techniques such as the one reported in [7] tends to design post processing based on morphological operations, which proved to be an effective towards removal of small holes inside the object regions. When part of the moving object is relatively still across a few frames, however, the edge maps in both (3) and (4) will fail to include those pixels, and thus create large holes inside the segmented object, for which the morphological operations will not be able to recover those missing parts inside or at the boundary of the segmented object. To this end, we propose a constrained region-growing technique to recover those missing parts. Unlike the normal region growing used by those spatial segmentation techniques for still images, our proposed region-grow is under certain constraints to reflect the fact that object segmentation has been done by change detections across adjacent frames. Therefore, the constraints include: (i) the seed selection is fixed at those edge points at the boundary of the final edge map; (ii) the number of pixels outside the first and the last edge points must be smaller than a certain limit. In other words, if the majority of the pixels on any row are outside the boundary of the edge map, the constrained region grows will not be applied.

Given the final edge map ME_n , we examine those pixels outside the first and the last pixel in each row to see if any further growing can be facilitated by using the pixel at the boundary of the edge map as seeds. Specifically, given the set of pixels outside the first and the last edge points in the i th row: $PO_i = \{PO_1, PO_2, \dots PO_k\}$, we decide whether the region of those edge points should be grown into any of the points inside PO_i or not by the following testing:

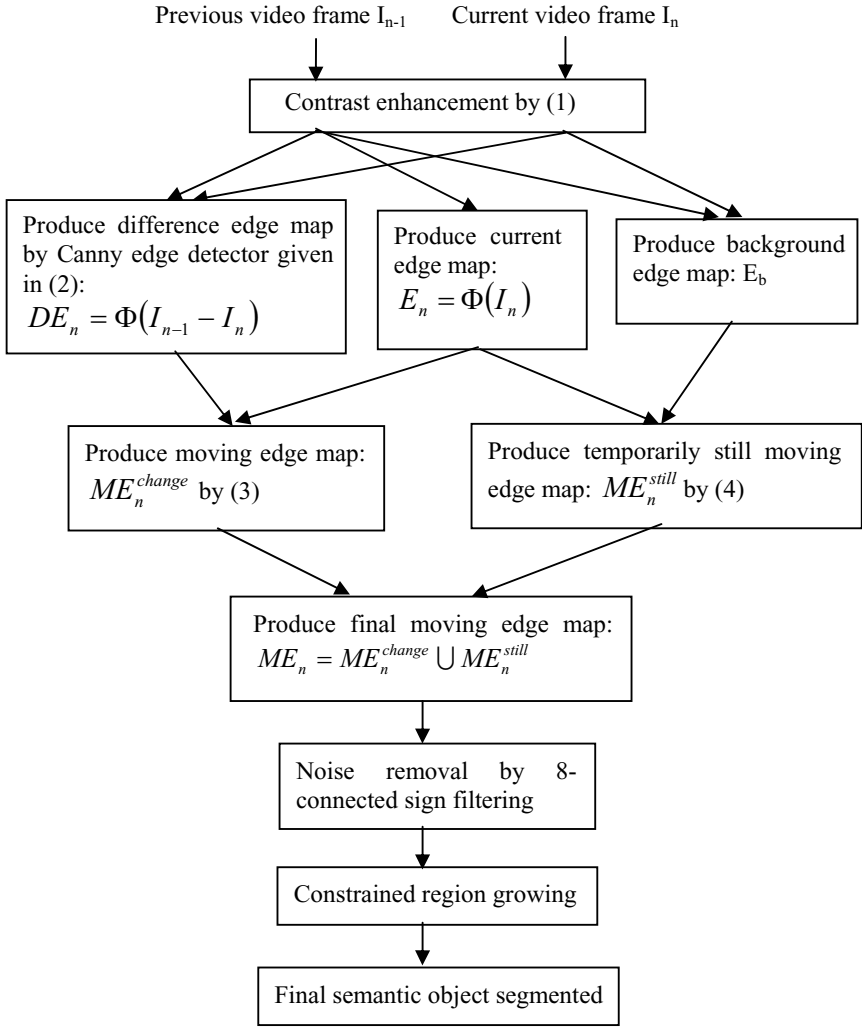


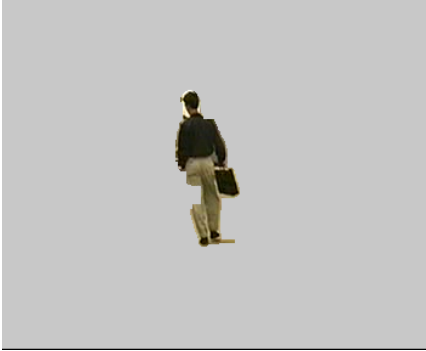
Fig. 1. Summary of the proposed algorithm

$$PO_i = \begin{cases} e_i & \text{if } \|PO_i - e\| \leq T_e \\ PO_i & \text{else} \end{cases} \quad (5)$$

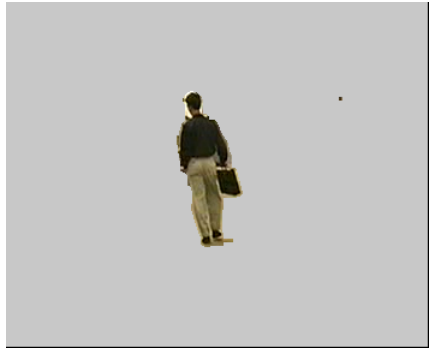
where T_e is a threshold indicating that the pixel tested is very similar to e , which is the first or the last edge pixel depending on which of these two edge points is closest to the position of PO_i . If the condition is satisfied, the PO_i will be grown into the edge points. Otherwise, they will stay as they are outside the edge map.

The above post processing will also apply to those points along the columns. After the post processing, the VO is extracted by logic AND operation of both row and column candidates as described in [7].

In summary, the proposed segmentation algorithm is highlighted by Figure 1.



(a): *segmentation by benchmark for Hall-monitor frame 71*



(b): *segmentation by the proposed for Hall-monitor frame 71*



(c): *segmentation by benchmark for Clair*



(d): *segmentation by the proposed for Clair*



(e): *segmentation by benchmark for Mother-daughter (frame304)*



(f): *segmentation by the proposed for Mother-daughter (frame 304)*

Fig. 2. Comparison of segmentation results between the benchmark and the proposed with linear transform given in (1)

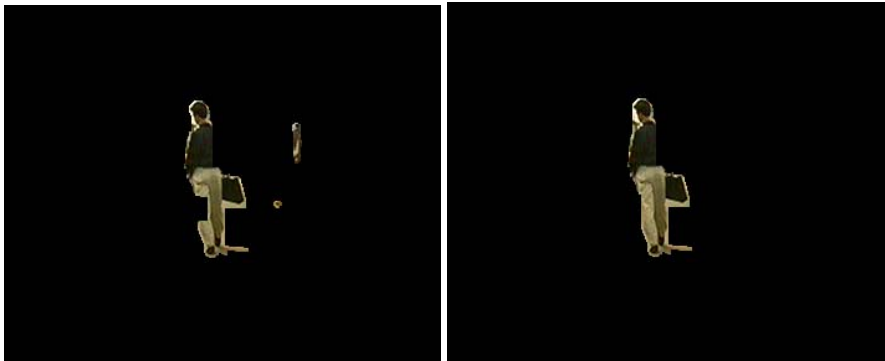
3 Experimental Results and Conclusions

To evaluate the proposed algorithm, we used a set of three video clips: *Hall-monitor*, *Claire*, and *Mother-daughter*, all of which are publicly available and *Hall-monitor* is the same as that used in [7]. In order to enable detailed analysis of how each element of the proposed algorithm actually contributes to the effect of final video object segmentation, we implemented the VO segmentation algorithm as in [7] as our benchmark, and carried out experiments each time one element of the proposed algorithm is added. These elements include: (i) linear transform for contrast enhanced edge detection; (ii) filter design for noise removal; and (iii) constrained region-growing.



(a): Segmentation result by the proposed for *Hall-monitor* (frame71) (b): Segmentation result by the proposed for *Mother-daughter* (frame304)

Fig. 3. Illustrations of segmentation by the proposed with both linear transform and noise removal filtering



(a): Segmentation by benchmark(frame 73)

(b): segmentation by the proposed (frame 73)

Fig. 4. Comparison of segmented results by benchmark and the proposed, where only the constraint region growing is considered

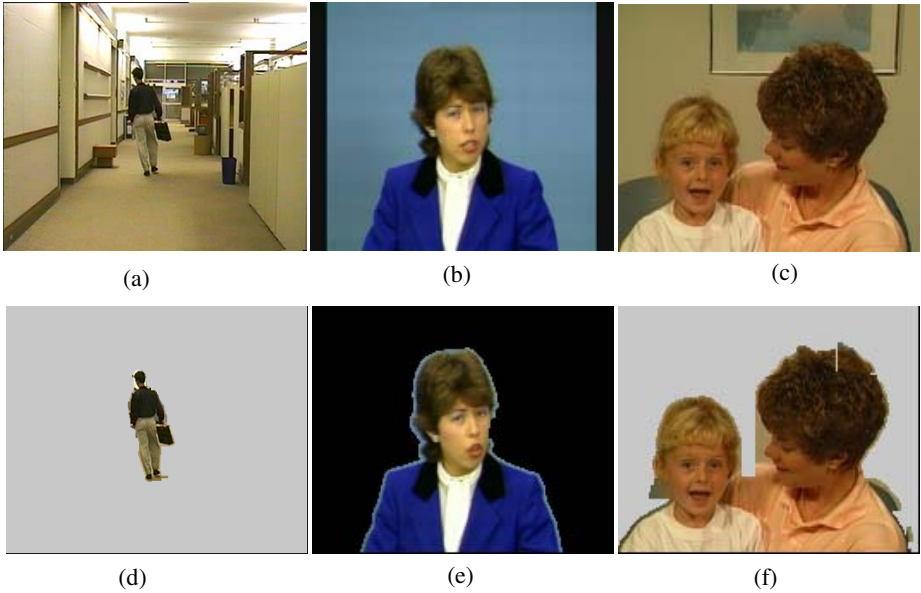


Fig. 5. Final segmentation results by the proposed algorithm: (a)-(c) are originals, and (d)-(e) are the segmented video objects

Figure 2 illustrates the comparison of the segmented results, where part-(a), (c), (e) and (f) are the segmented objects by the benchmark and part-(b), (d), (f) and (g) are the segmented objects by the proposed algorithm, where only the linear transform is included. As seen, the proposed linear transform introduced additional noise while the segmentation accuracy is improved.

Figure-3 illustrates the segmented results by the proposed algorithm, where both linear transform and the noise removal filter are considered, from which it can be seen that the noise introduced is effectively removed.

Figure-4 illustrates the comparison of segmented results between the benchmark and the proposed, where part-(a) and (c) represents the results of the benchmark, and part-(b) and (d) the results of the proposed with only the element of constrained region-growing. Although the proposed constrained region-growing can not recover all the missing parts, it can still be seen that the proposed algorithm does recover the missing part inside the left leg, which has achieved significant improvement compared with the benchmark.

Finally, by gathering all the elements, the full segmented video objects by the proposed algorithm can be illustrated in Figure 5. Note that all the figures illustrated here are much larger than those given in references [6-15]. If we make the pictures smaller, the segmentation results will look better as those boundaries will look smoother.

In this paper, we proposed an automatic semantic object segmentation scheme to provide a possible solution for the under-segmentation problem experienced by most existing segmentation techniques [6-15]. From the experimental results shown in Figure 2 to Figure 5, it can be seen that, while the proposed algorithm can effectively recover those missing parts inside the video object, it inevitably introduces some of the back-

ground points into the object region, which can be referred to as over-segmentation, for which further research is being organized around: (i) looking for other cues to provide additional semantic information for segmentation; (ii) combinational approaches in both change detection and background registration[9]; and (iii) inclusion of further spatial segmentation elements such as snake modeling, and water shed [1-5] etc.

Finally, the authors wish to acknowledge the financial support from the Chinese Academy of Sciences and European Framework-6 IST programme under the IP project: Live staging of media events (IST-4-027312).

References

1. K. Harris S. N. Efstratiadis, N. Maglaveras, and A. K. Katsaggelos, "Hybrid image segmentation using watersheds and fast region merging," *IEEE Trans. on image processing*, Vol.7, No.12, pp. 1684-1699, 1998
2. L. Vincent, P. Soille, "Watersheds in digital spaces: an efficient algorithm based on immersion simulations," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol.13, Issue 6, pp. 583-598, 1991
3. R. J. O'Callaghan, and D. R. Bull, "Combined Morphological-Spectral unsupervised image segmentation," *IEEE Trans. on image processing*, Vol.14, No. 1, pp. 49-62, 2005
4. Tsai YP, Lai C.C. Hung YP et al. A Bayesian approach to video object segmentation via merging 3-D watershed volumes *IEEE Trans. On Circuits and Systems for Video Tech.* 15 (1) pp. 175-180 Jan. 2005
5. Jung CR, Scharcanski J 'Robust watershed segmentation using wavelets' *Image and Vision Computing* 23 (7): 661-669 Jul. 1 2005
6. Kim M. et al. 'A VOP generation tool: automatic segmentation of moving objects in image sequences based on spatio-temporal information', *IEEE Trans. Circuits, Systems for Video Technology*, Vol 9, No 8, 1999, pp. 1216-1227
7. Kim C. and Hwang J.N. 'fast and automatic video object segmentation and tracking for content-based applications', *IEEE Trans. Circuits and Systems for Video Technology*, Vol 12, No. 2, 2002, pp. 122-128
8. Salembier P. and Pardas M. 'Hierarchical morphological segmentation for image sequence coding', *IEEE Trans. Image Processing*, Vol 3, No 5, 1994, pp. 639-648
9. Chien S.Y. et. Al. 'Efficient moving object segmentation algorithm using background registration technique', *IEEE Trans. Circuits and Systems for Video Technology*, Vol 12, No 7, 2002, pp. 577-589
10. Shamim A. and Robinson A. 'Object-based video coding by global-to-local motion segmentation', *IEEE Trans. Circuits, Systems for Video Technology*, Vol 12, No 12, 2002, pp. 1106-1115
11. Feng G.C. and Jiang J "Image segmentation in compressed domain" *Journal of Electronic Imaging*, Vol .12, No 3, SPIE, 2003, pp. 390-397
12. Toklu C. et. Al. 'Semi-automatic video object segmentation in the presence of occlusion', *IEEE Trans. Circuits and Systems for Video Technology*, Vol 10, No 4, 2000, pp. 624-635
13. Kervrann C. and Heitz F. 'Statistical deformable model-based segmentation of image motion', *IEEE Trans Image Processing*, Vol 8, No 4, 1999, pp. 583-594
14. Meier T. and Ngan K. 'Automatic segmentation of moving objects for video object plane generation', *IEEE Trans. Circuits, Systems for Video Tech.*, Vol 8, No 5, 1998, p. 525
15. Xu Y. et. Al. 'Object-based image labeling through learning by example and multi-level segmentation', *Pattern Recognition*, Vol 36, pp. 1407-1423, 2003

Spatio-temporal Composite-Features for Motion Analysis and Segmentation

Raquel Dosal¹, Xosé M. Pardo¹, Xosé R. Fdez-Vidal², and Antón García¹

¹ Dept. Electrónica e Computación, Universidade de Santiago de Compostela, Campus Universitario Sur, s/n, 15782, Santiago de Compostela, A Coruna, Spain
rdosal@usc.es, pardo@dec.usc.es, antongd@usc.es

² Escola Politécnica Superior, Univ. de Santiago de Compostela, Campus Universitario, s/n, 27002, Lugo, Spain
faxose@usc.es

Abstract. Motion estimation by means of spatio-temporal energy filters –velocity tuned filters– is known to be robust to noise and aliasing and to allow an easy treatment of the aperture problem. In this paper we propose a motion representation based on the composition of spatio-temporal energy features, i.e., responses of a set of filters in phase quadrature tuned to different scales and orientations. Complex motion patterns are identified by unsupervised cluster analysis of energy features. The integration criterion reflects the degree of alignment of maxima of the features’s amplitude, which is related to phase congruence. The composite-feature representation has been applied to motion segmentation with a geodesic active model both for initialization and image potential definition. We will show that the resulting method is able to handle typical problems, such as partial and total occlusions, large inter-frame displacements, moving background and noise.

1 Introduction

In this paper we deal with the problem of segmentation of apparent-motion. Apparent-motion segmentation can be stated as the identification and classification of regions undergoing the same motion pattern along a video sequence. Usually, segmentation is based on some low level feature describing the motion of each pixel in a video frame. So far, the variety of approaches to deal with the problems of motion feature extraction and motion segmentation that has been proposed in literature is huge. However, all of them suffer from different shortcomings and up to date there is no completely satisfactory solution. Segmentation and tracking techniques developed so far present diverse kinds of problems: restriction to some specific motion model, as methods those based in the Hough transform [1]; inability to deal with occlusions and abrupt changes, like Kalman filter based approaches [2]; need of some prior model or template [3]; sensitivity to noise and lack of correlation among segmentations from different frames, as it happens with Bayesian classification methods [4].

Many other problems are a consequence of the low-level motion representation underlying the segmentation model. It usually involves the estimation of the

temporal derivative of the sequence. When used directly to detect mobile points [5], this measure is limited to static background and can not discriminate independent motion patterns. Most representations use the inter-frame difference to estimate optical flow. Optical flow estimation algorithms present diverse kinds of problems as well [6]. In general, they assume brightness constancy along frames, which in real situations does not always hold, and restrict allowed motions to some specific model. Particularly, differential methods are not very robust to noise, aliasing, occlusions and large inter-frame displacements and present the aperture problem, i.e., they operate locally so they do not yield reliable estimations of the direction of motion. Matching techniques require prior knowledge about the tracked object and, generally, consider only rigid transformations.

Alternatively, energy based algorithms [7,8,9,10,11] do not employ the temporal derivative, but estimate motion from the responses of spatio-temporal filter pairs in quadrature, tuned to different scales and orientations, which is translated into sensitivity to 2D orientation, speed and direction of movement. These techniques are known to be robust to noise and aliasing, to give confident measurements of velocity and to allow an easy treatment of the aperture problem. This representation is further developed in [12], by defining spatio-temporal *integral* energy features as clusters of elementary velocity-tuned filters. The integration criterion, which introduces no prior information, is based on the congruence or similitude of a set of local statistics of the energy features.

In this work we present a motion representation scheme based on the clustering of non-causal energy filters that introduces a new integration criterion, that improves the computational cost and performance of earlier approaches. It is inspired on the importance of Phase Congruence (PC) [13] and local energy maxima [14] as low level cues of information for the HVS. Our integration criterion to group band-pass features reflects the degree of concurrence of their energy maxima. We estimate the degree of alignment of energy maxima of pairs of band-pass features from the correlation between their energy maps. Moreover, spatio-temporal filters with rotational symmetry will be introduced. We will show this representation handles motion patterns composed of different speeds, directions and scales, distinguishes visually independent patterns and is robust to noise, moving background, occlusions and large displacements.

The representation model will be applied here to segmentation in combination with a geodesic active model [15]. Composite motion features will be applied directly, avoiding the estimation of optical flow. They will be employed for both the definition of the image potential and for the initialization of the model in each frame. This solution is novel in the sense that both initialization and segmentation is based on motion information. In [5], initialization is defined from the segmentation from previous frame, which is problematic in case of occlusions. In [16], motion features are employed for initialization in each frame while segmentation is based only in spatial information.

The outline of this paper is as follows. Section 2 is dedicated to the composite feature representation model. Section 3 is devoted to the proposed method for segmentation with active models. In section 4 we illustrate the behavior

the model in different problematic situations, including some standard video sequences. In 5 we expound the conclusions extracted from this work and set out some lines for future work.

2 Detection of Composite-Features

Identification of composite energy features involves the decomposition of the sequence into a set of band-pass features and their subsequent grouping according to some dissimilarity measure. Composite-features are then reconstructed as a combination of the band-pass features in each clusters. The following subsections detail the process.

2.1 Bank of Spatio-temporal Energy Filters

As mentioned earlier, we employ non-causal band-pass features, This allows treating the temporal dimension as a third spatial dimension. In particular, here we apply an extension to 3D [17] of the log Gabor function [18] that presents rotational symmetry. The filter is designed in the frequency domain. The filters' transfer function T is designed in spherical frequency coordinates (ρ, ϕ, θ) .

$$T_i(\rho, \phi, \theta) = \exp\left(-\frac{(\log(\rho/\rho_i))^2}{2(\log(\sigma_{\rho i}/\rho_i))^2} - \frac{\alpha_i(\phi, \theta)^2}{2\sigma_{\alpha i}^2}\right) \quad (1)$$

with,

$$\alpha_i(\phi, \theta) = \arccos\left(\frac{\mathbf{f} \cdot \mathbf{v}_i}{\|\mathbf{f}\|}\right) \quad (2)$$

where $(\rho_i, \phi_i, \theta_i)$ is the central frequency of the filter, $\sigma_{\rho i}$ and $\sigma_{\alpha i}$ are the standard deviations, $\mathbf{v}_i = (\cos(\phi_i)\cos(\theta_i), \cos(\phi_i)\sin(\theta_i), \sin(\theta_i))$ and \mathbf{f} is the position vector of a point in the frequency domain, expressed in Cartesian coordinates. It is the product of a radial term, the log Gabor function, selective in frequency and an angular term, a gaussian in the angular distance α [19], which provides selectivity in spatial orientation, direction or motion and speed.

The filter bank is composed of a predefined set of the previous energy filters with $\lambda_i = 2/\rho_i = \{4, 8, 16, 32\}$, θ_i is sampled uniformly and ϕ_i is sampled to produce constant *density* with θ by forcing equal arc-length between adjacent ϕ_i samples over the unit radius sphere. Following this criterion, the filter bank has been designed using 23 directions, i.e. (ϕ_i, θ_i) pairs, yielding 92 bands. $\sigma_{\rho i}$ is calculated for each band to obtain 2 octave bandwidth and $\sigma_{\alpha i}$ is set to 25° for all orientations.

From the previous bank, only *active* channels, contributing the most to energy, are selected as those comprising some value of $E = \log(|F| + 1)$, where F is the Fourier transform, over its estimated maximum noise level. Noise is measured in the band $\lambda < 4$ [20], as $\eta + 3\sigma$, where η is the average and σ is the standard deviation. Thresholding is followed by a radial median filtering, [17] to eliminate spurious energy peaks. Complex-valued responses ψ to active filters play the role of elementary energy features.

2.2 Energy Feature Clustering

Integration of elementary features is tackled in a global fashion, not locally (point-wise). Besides computational efficiency, this provides robustness since it intrinsically correlates same-pattern locations –in space and time–, avoiding re-covering of disconnected regions.

Our criterion for integration of frequency features is derived from the concept of Phase Congruence (PC), i.e., the local degree of alignment on the phase of Fourier components. Features where Fourier components are locally in phase play an important role in biological vision [13]. Our grouping strategy consists on finding the frequency bands that are contributing to a point or region of locally maximal PC. Since points of locally maximal PC are also points of locally maximal energy density [14], we will use local energy as a detector of relevant features. Then, subband images contributing to the same visual pattern should present a large degree of alignment between their local energy maxima. The dissimilarity between two subband features is determined by estimating the degree of alignment between the local maxima of their local energy. Alignment is quantified using the correlation coefficient r of the energy maps of each pair $\{\psi_i, \psi_j\}$ of subband features. If $A = \|\psi\| = (\Re(\psi)^2 + \Im(\psi)^2)^{1/2}$, the actual distance is calculated from $r(A_i, A_j)$ as follows

$$D_\rho(A_i, A_j) = \left(1 - \sqrt{(1 - r(A_i, A_j)) / 2}\right)^2 \tag{3}$$

To determine the clusters from the dissimilarities, hierarchical clustering has been applied, using a Ward’s algorithm to determine inter-cluster distance, which has proved to improve other metrics [21]. To determine the number of clusters N_c , the algorithm is run for each possible N_c and the quality of each resulting partition is evaluated according to a validity index, the modified Davies-Boulding index [22].

2.3 Composite-Feature Reconstruction

The response ψ of an energy filter is a complex-valued sequence. Its real and imaginary components account for even and odd symmetric features respectively. The use of $\Re(\psi)$, $\Im(\psi)$ or $A(\psi)$ in the construction the composite-feature Ψ will depend on its purpose. Here we define the general rule for the reconstruction of Ψ based on a given representation E of the responses of the filters.

$$\Psi^j(x, y, t) = \frac{\sum_{i \in \Omega_j} \tilde{E}_i}{\text{Card}(\Omega_j)} \sum_{i \in \Omega_j} E_i(x, y, t) \tag{4}$$

where Ω_j is the set of all features in cluster j and \tilde{E}_i results from applying a sigmoidal thresholding to feature E_i . The first factor in the right side of previous equation weights the response in a point favoring locations with a large fraction of features contributing to it, avoiding artifacts caused by individual elementary features.

For visualization purposes, we use the even-symmetric representation of the composite feature $\Psi_{even}^j = \Psi^j(E_i = \Re(\psi_i))$. The full-wave rectified odd-symmetric representation $\Psi_{odd}^j = |\Psi^j(E_i = \Im(\psi_i))|$ is employed in the definition of image potentials to reflect motion contours. Initialization is determined from the amplitude representation $\Psi_{amp}^j = \Psi^j(E_i = \|\psi_i\|)$, except for objects with uniform contrast, where $\max(\pm\Psi_{even}^j, 0)$ is used, with sign depending on the specific contrast sign.

3 Motion Pattern Segmentation

In this section we prove the usefulness of the developed representation in segmentation and tracking. The chosen segmentation technique is the geodesic active model as implemented in [15], due to its ability to introduce continuity and smoothness constraints, deal with topological changes and simultaneously detect inner and outer regions. Our segmentation model involves the selection, by user interaction, of one of the identified composite-features Ψ . From that pattern, we derive the initial state and the image potential of the model in each frame. Segmentation of a frame is illustrated in figure 1.

3.1 Geodesic Active Model

The geodesic active model represents a contour as the zero-level set of a distance function u . The evolution of the level-set determines the evolution of the contour. Let $\Omega := [0, a_x] \times [0, a_y]$ be the frame domain and consider a scalar image $u_0(x, y)$ on Ω . We employ here symbol τ for time in the evolution equations of u and t for the frame index. Then, the equations governing the evolution of the implicit function are the following.

$$u(x, y, t = t_k, \tau = 0) = u_0(x, y, t = t_k) \text{ on } \Omega \tag{5}$$

$$\frac{\partial u}{\partial t} = g(s)|\nabla u| \cdot (\kappa + c) + \nabla g(s) \cdot \nabla u \text{ on } \Omega \times (0, \infty) \tag{6}$$

$$g(s) = (1 + (s/s_{min})^2)^{-1} \tag{7}$$

where s_{min} and c are real constants, s is the selected image feature and κ is the curvature. The role of the curvature can be interpreted as a geometry dependent velocity. Constant c represents a constant velocity or *advection* velocity. Here, $c = 0$. The factor $g(s)$ is called the image potential, which decreases in the presence of image features. In the first term, it modulates velocity ($c + \kappa$) so as to stop the evolution of the contour at the desired feature locations. The second term is the image dependent term, which pushes the level-set towards image features. s_{min} plays the role of a feature threshold.

The image feature s is determined from Ψ_{odd} . A dependence on pure spatial features, namely, a spatial contour detector, is also introduced to close the contour when part of the boundary of the moving object remains static –when partial occlusion or when part of the moving contour is parallel to the direction

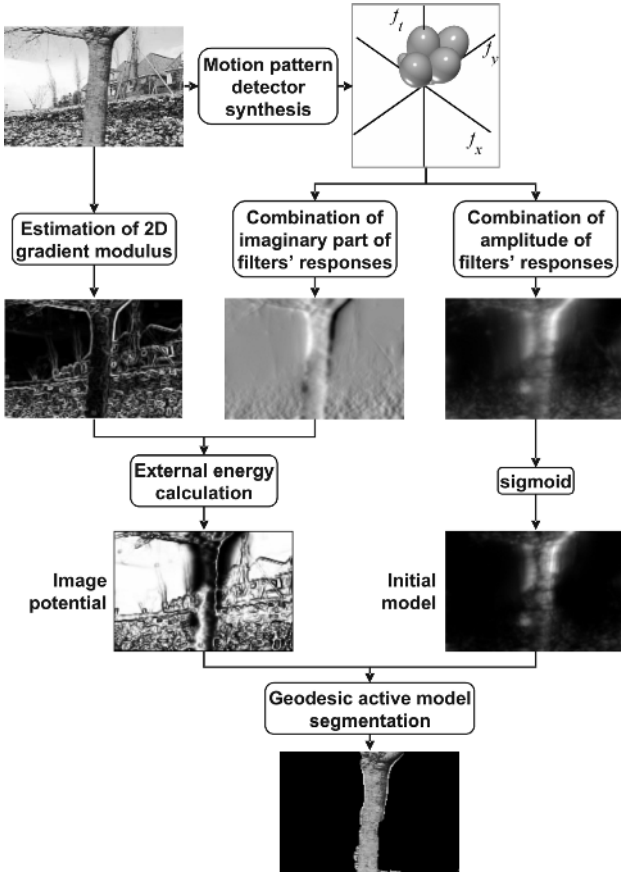


Fig. 1. Scheme of the segmentation technique. Composite feature detector is represented by the isosurfaces of the constituent filters with $T_i = \exp(-1/2)$.

of motion. Accordingly, the image feature s is a weighted sum of a spatial term C_s and a motion term C_m

$$\begin{aligned}
 s &= w_s C_s + s_m C_m, \quad \text{with } w_s = 1 - w_m \text{ and } 0 \leq w_m \leq 1, \\
 C_s(x, y, t_k) &= \|\nabla I^*(x, y, t_k)\| / \max(\|\nabla I^*(x, y, t_k)\|), \\
 C_m(x, y, t_k) &= \frac{1}{1 + \exp\{-K(C_s(x, y, t_k) - C_0)\}} \frac{\Psi_{\text{odd}}(s, y, t_k)}{\max(\Psi_{\text{odd}}(x, y, t_k))},
 \end{aligned} \tag{8}$$

where C_0 and K are a positive real constants and asterisk indicates regularization. The first factor in the expression for C_m is a sigmoidal thresholding of the spatial contour detector C_s that modulates the contribution of the motion feature depending on the concurrent presence of a spatial gradient. This is to minimize the effect of temporal diffusion in Ψ . Regularization of a frame is accomplished here by feature-preserving 2D anisotropic diffusion, which brakes diffusion in the

presence of contours and corners [23]. The specific values of parameters involved are $C_0 = 0.1$, $K = 20$, $w_m = 0.9$. In equation (7), s_{min} is calculated so that, on average, $g(s(x, y)) = 0.01$, $\forall x, y : C_m(x, y) > 0.1$. Considering the geodesic active model in a front propagation framework, $g = 0.01$ means a sufficiently slow speed of the propagating front to produce stopping in practical situations.

3.2 Initialization

The initial state of the geodesic active model is defined, in the general situation, from Ψ_{amp} unless other solution is specified. To enhance the response of the cluster we apply a sigmoid thresholding and remapping to the interval $[-1, 1]$. To handle the situation where the object remains static during some frames, the initial model is defined as the weighted sum of two terms, one associated to the motion pattern and the other to the segmentation in previous frame. The zero-level of the resulting image is the initial state of the contour.

$$u_0(x, y, t_k) = w_k (2 / (1 + \exp(-K(\Psi_{amp}(x, y, t_k) - \Psi_0))) - 1) + w_{k-1} u(x, y, t_{k-1}, \tau_{max}), \text{ with } w_{k-1} = 1 - w_k, \quad 0 \leq w_k \leq 1 \quad (9)$$

In the experiments presented in next section, $w_k = 0.9$, $K = 20$ and $\Psi_0 = 0.1$.

4 Results

In this section, some results are presented to show the behavior of the method in problematic situations. The video sequences and corresponding segmentations are available at [24]. The results are compared to an alternative implementation similar to that in [5]: the initial state is the segmentation of the previous frame and the image potential depends on the inter-frame difference. However, instead of defining the image potential from the temporal derivative using a Bayesian classification, the image potential is the same as with our method, except that the odd-symmetric representation of the motion pattern is replaced by the inter-frame difference $I_t(x, y, t_k) = I(x, y, t_k) - I(x, y, t_{k-1})$. This is to compare the performance of our low-level features with inter-frame difference under the equal conditions. The initial state for the first frame is defined by user interaction.

4.1 Moving Background

This example is part of the the well-known "flower garden" sequence, a static scene recorded by a moving camera. The inter-frame difference detects motion at every image contour, producing deep minima in the image potential all over the image. Therefore, the active model is not able to distinguish foreground objects from background –see figure 2. It is not possible to isolate different motion patterns by simple classification of I_t values. In contrast, visual pattern decomposition allows isolation of different motion patterns with different speeds. The image potential in our approach considers only the motion pattern corresponding to the foreground, leading to a correct segmentation –see figure 1.

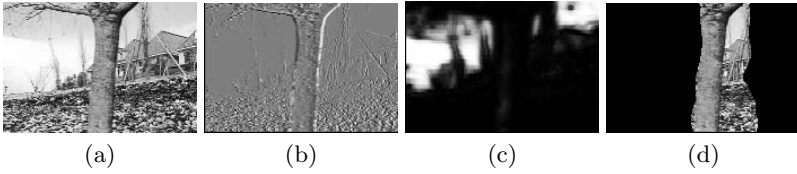


Fig. 2. A frame of the "flower garden" sequence. (a) Input data. (b) Inter-frame difference. (c) Image potential from I_t . (d) Segmentation using I_t based active model.

4.2 Large Inter-frame Displacements

When the sampling rate is too small in relation to the speed of the moving object, it is difficult to find the correspondence between the positions of the object in two consecutive frames. Most optical flow estimation techniques present strong limitations in the allowed displacements. Differential methods try to find the position of a pixel in the next frame imposing some motion model. Frequently, the search is restricted to a small neighborhood. This limitation can be overcome by coarse-to-fine analysis or by imposing smoothness constraints [6]. Still, large displacements are usually problematic. The Kalman filter is not robust to abrupt changes when no template is available [2]. When using the inter-frame difference in combination with an active model and initialization with previous frame, the correspondence is yielded by the evolution of the model from the previous state to the next one [5]. However, if the previous segmentation does not intersect the object in the following frame, the model is not able to track the target, as shown in the example in figure 3.

When using energy features, the composite motion patterns are isolated from each other. In this way, the correspondence of the motion estimations in different frames is naturally provided by the representation scheme, as shown in the right column of figure 3. This is also a property of techniques based on the Hough transform [1] but they are limited to constant speed and direction motion. In this example there is an oscillating movement, so it does not describe a line or a plane in the velocity-space. Unlike the Hough transform, composite features combine elementary velocity-tuned features to deal with complex motion patterns. The ability to represent complex pattern in a global fashion, leads the active model to a correct segmentation of the ball –see figure 3, third column– besides the large displacement produced and the changing direction or movement.

4.3 Occlusions

In the sequence in first column of figure 4, the mobile object is totally occluded during several frames. This is a severe problem for many tracking algorithms. In region classification techniques [4], the statistical models extracted for each region can be employed for tracking by finding the correspondence among them in different frames. This is not straightforward when the object disappears and reappears in the scene. The same problem applies for Kalman filtering [2]. In the

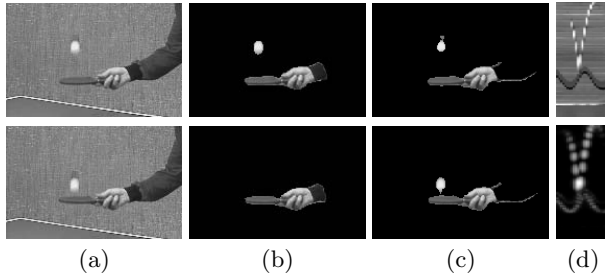


Fig. 3. (a) Two consecutive frames of the "table tennis" sequence. (b) Segmentations of frames in (a) using the alternative. (c) Segmentations of frames in (a) using the composite-feature active model. (c) $x - t$ cut plane of the input data (*top*) and Ψ_{amp} (*bottom*).

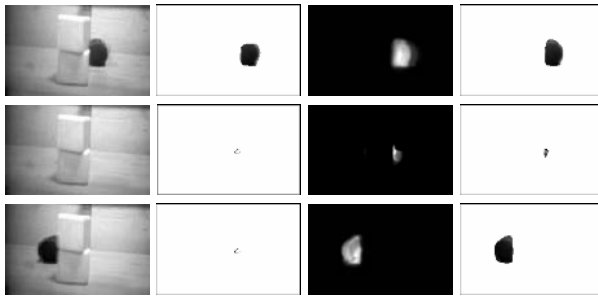


Fig. 4. Three frames of a sequence showing total occlusion. *Left:* Input data. 2^{nd} *Column:* Segmentation using the alternative active model. 3^{rd} *Column:* Initial model from $\max(-\Psi_{even}, 0)$ *Right:* Segmentation using the composite-feature active model.

case of active models with initialization with previous frame is applied, when total occlusions, the model collapses and no initial model is available in subsequent frames, as can be observed in second column of figure 4. A new stage of motion detection can be used to reinitialize the model [5], but it can not be ensured that the newly detected motion feature corresponds to the same pattern.

With our representation this problem is automatically solved, as in the case of large inter-frame displacements. In third column of figure 4, we observe that, when the object reappears, initialization is automatically provided by the motion pattern without extra computation and without the need of a prior model, leading to a correct segmentation –see figure figure 4, right column. In this example the initial model is defined using $\max(-\Psi_{even}, 0)$.

Figure 5 shows another example presenting occlusions where the occluding object is also mobile. As can be seen, the alterative active model fails in segmenting both motion patterns, both due to initialization with previous segmentation and incapability of distinguishing both motion patterns, while our model properly segments both patterns using the composite-features provided by our representation scheme.

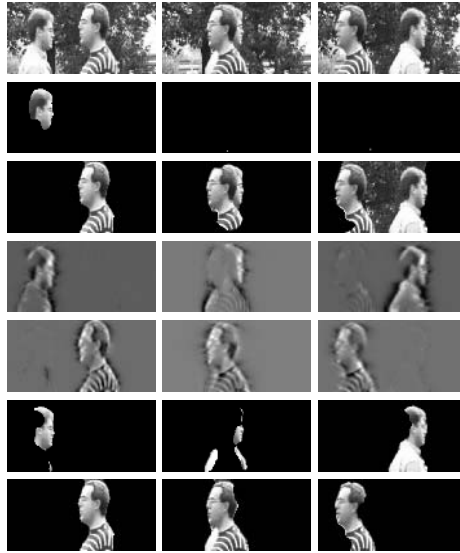


Fig. 5. Three frames of a sequence showing two occluding motion patterns. *1st Row:* Input data. *2nd and 3rd Rows:* Inter-frame based segmentation, using a different initialization for each of the motion patterns. *4th and 5th Rows:* Ψ_{even} of two of the obtained composite-features, corresponding to the two motion patterns. *6th and 7th Rows:* Segmentations produced using composite-features from rows 4th and 5th respectively.



Fig. 6. Two frames of the "silent" sequence: *Left:* Input data. *2nd Column:* Segmentation using the alternative active model. *3rd Column:* $\max(\Psi_{even}, 0)$ of the selected motion pattern. *Right:* Segmentation using the composite-feature active model.

4.4 Complex Motion Patterns

The following sequence, a fragment of the standard movie known as "silent" is relatively complex, presenting different moving parts, each one with variable speed and direction and deformations too, over a textured static background. As can be observed in left column of figure 6, the motion pattern of the hand cannot be properly described by an affine transformation. Moreover, the brightness constancy assumption is not verified here. The active model based on the inter-

frame difference is not able to properly converge to the contour of the hand –figure 6, second column– due to both interference of other moving parts or shadows and wrong initialization. The composite-feature representation model is able to isolate the hand and properly represent its changing shape in different frames –figure 6, third column. Hence, the initial state of the active model is already very close to the object’s shape, thus ensuring and accelerating convergence –figure 6, right column.

5 Conclusions

This work proposes a new motion representation scheme for video sequences based on non-causal composite energy features. It involves clustering of elementary band-pass features under a criterion of spatio-temporal phase congruence. This motion representation is able to isolate visually independent motion patterns without using priori knowledge. The model intrinsically correlates information from different frames, providing it with robustness to occlusions and large inter-frame displacements. Furthermore, it is suitable for complex motion models thanks to the composition of elementary motion features.

Our representation model has been applied to the definition of image potential and initialization of a geodesic active model for segmentation and tracking. The proposed technique presents good performance in many of the typical problematic situations, such as noise, moving background, occlusions, and large inter-frame displacements, while previous solutions do not deal with all these issues. In the comparison with an alternative implementation, that employs segmentation of previous frame for initialization and inter-frame difference for definition of image potential, our method shows enhanced behavior.

Acknowledgements

This work has been financially supported by the Xunta de Galicia through the research project PGIDIT04TIC206005PR.

References

1. Sato, K., Aggarwal, J.: Temporal spatio-temporal transform and its application to tracking and interaction. *Comput. Vis. Image Underst.* **96** (2000) 100–128
2. Boykov, Y., Huttenlocher, D.: Adaptive bayesian recognition in tracking rigid objects. In: *IEEE CVPR*, Vol. 2. (2000) 697–704
3. Nguyen, H., Smeulders, A.: Fast occluded object tracking by a robust appearance filter. *IEEE Trans. Pattern Anal. Mach. Intell.* **26** (2004) 1099–1104
4. Montoliu, R., Pla, F.: An iterative region-growing algorithm for motion segmentation and estimation. *Int. J. Intell. Syst.* **20** (2005) 577–590
5. Paragios, N., Deriche, R.: Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** (2000) 266–279

6. Barron, J., Fleet, D., Beauchemin, S.: Performance of optical flow techniques. *Int. J. Comput. Vis.* **12** (1994) 43–77
7. Heeger, D.: Model for the extraction of image flow. *J. Opt. Soc. Am. A* **4** (1987) 1555–1471
8. Simoncelli, E., Adelson, E.: Computing optical flow distributions using spatio-temporal filters. Technical Report 165, MIT Media Lab. Vision and Modeling, Massachusetts (1991)
9. Watson, A., Ahumada, A.: Model for human visual-motion sensing. *J. Opt. Soc. Am. A* **2** (1985) 322–342
10. Adelson, E., Bergen, J.: Spatiotemporal energy models for the perception of motion. *J. Opt. Soc. Am. A* **2** (1985) 284–299
11. Fleet, D.: *Measurement of Image Velocity*. Kluwer Academic Publishers, Massachusetts (1992)
12. Chamorro-Martínez, J., Fdez-Valdivia, J., García, J., Martínez-Baena, J.: A frequency domain approach for the extraction of motion patterns. In: *IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, Vol. 3. (2003) 165–168
13. Morrone, M., Owens, R.: Feature detection from local energy. *Pattern Recognit. Lett.* **6** (1987) 303–313
14. Venkatesh, S., Owens, R.: On the classification of image features. *Pattern Recognit. Lett.* **11** (1990) 339–349
15. Weickert, J., Kühne, G.: Fast methods for implicit active contour models. In Osher, S., Paragios, N., eds.: *Geometric Level Set Methods in Imaging, Vision and Graphics*. Springer, New York (2003) 43–58
16. Tsechpenakis, G., Rapantzikos, K., Tsapatsoulis, N., Kollias, S.: A snake model for object tracking in natural sequences. *Signal Process.-Image Commun.* **19** (2004) 219–238
17. Dosil, R., Pardo, X., Fdez-Vidal, X.: Data driven detection of composite feature detectors for 3D image analysis. *Image Vis. Comput.* **32** (2006) 225–238
18. Field, D.: What is the goal of sensory coding. *Neural Comput.* **6** (1994) 559–601
19. Faas, F., van Vliet, L.: 3D-orientation space; filters and sampling. In Bigun, J., Gustavsson, T., eds.: *LNCS 2749: Scandinavian Conference on Image Analysis*, Berlin Heidelberg, Springer-Verlag (2003) 36–42
20. Kovese, P.: *Invariant Measures of Image Features from Phase Information*. PhD thesis, The University of Western Australia (1996)
21. Jain, A., Dubes, R.: *Algorithms for Clustering Data*. Prentice Hall (1988)
22. Pal, N., Biswas, J.: Cluster validation using graph theoretic concepts. *Pattern Recognit.* **30** (1996) 847–857
23. Dosil, R., Pardo, X.: Generalized ellipsoids and anisotropic filtering for segmentation improvement in 3D medical imaging. *Image Vis. Comput.* **21** (2003) 325–343
24. URL: http://www-gva.dec.usc.es/~rdosil/motion_segmentation_examples.htm (2006)

New Intra Luma Prediction Mode in H.264/AVC Using Collocated Weighted Chroma Pixel Value

Ik-Hwan Cho, Jung-Ho Lee, Woong-Ho Lee, and Dong-Seok Jeong

Department of Electronic Engineering, Inha University, Yonghyun-Dong, Nam-Gu, Incheon,
Republic of Korea
{teddydino, jungho, ltleee}@inhaian.net, dsjeong@inha.ac.kr

Abstract. Intra coding in current hybrid video coding method has very important functionality like low delay in decoder, random access and error resilience. Unfortunately coding efficiency of intra frame is very low relative to inter frame coding because of mismatch between current block and its predicted block. In this paper, new intra luma prediction algorithm which improves intra coding efficiency is proposed. The proposed additional intra luma prediction mode uses collocated chroma pixels and weight values to estimate correct spatial pattern of coded block. From neighboring blocks, weight value between chroma and luma values is calculated and then the predicted luma block is obtained by multiplying calculated weight value and collocated upsampled chroma block. The proposed method is effective for complex or non-directional macroblocks and experimental results show that the efficiency of intra coding is increased up to 0.6 dB.

1 Introduction

H.264/AVC [1] based on hybrid video coding algorithms is the new standard focusing on high coding efficiency, error resilience and network adaptability. For high coding efficiency, H.264/AVC uses several new tools such as new intra prediction, multiple reference frames, new integer transform and Context Adaptive Binary Arithmetic Coding (CABAC). New intra prediction method in H.264/AVC is one of the best improvements factors in H.264/AVC relative to the previous video coding standards.

In intra block coding the predicted block is obtained from neighboring blocks which are already encoded and reconstructed. And the predicted block is subtracted from current original block to obtain residual block. In general cases, intra prediction is applied to 4x4 and 16x16 luma blocks and 8x8 chroma blocks. There are 9 prediction modes for each 4x4 luma block, 4 modes for 16x16 luma block and 4 modes 8x8 chroma block totally. Encoder selects optimal prediction mode for each luma and chroma components to get minimum number of bits for coding residual block. Prediction algorithms in each component are basically same, even if block size and the number of modes to be used are different. In 4x4 luma block, 13 neighboring pixels are used for intra prediction. Fig. 1(a) shows example labeling for 4x4 luma block to be encoded.

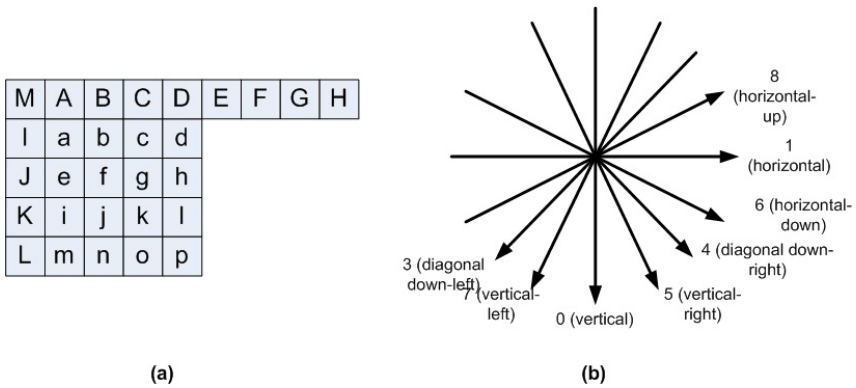


Fig. 1. Intra prediction mode. (a) Labeling of samples for 4x4 intra luma prediction. a, b, c,..., o and p represent current block samples and A, B, C,..., L and M represent neighboring samples to be used in prediction. (b) Directions of intra prediction mode for 4x4 luma block.

Intra prediction in H.264/AVC is based on block directionality and its correlation with neighboring blocks. All prediction modes except DC mode mean direction of prediction and Fig. 1(b) shows the directions for intra prediction. In intra prediction process, neighboring samples (capital letters in Fig. 1(a)) are extrapolated into current block sample position in according to each prediction mode. Using these intra prediction modes, H.264/AVC can show high coding efficiency relative to previous image and video coding standards such as JPEG-2000 [2], MPEG-2 Video [3] and MPEG-4 Visual [4].

New intra prediction scheme has many advantages rather than inter prediction such as low-delay, random access and error resilience. Despite of several good advantages, fatal disadvantage of intra coding relative to inter coding is its low coding efficiency because of mismatch between original and the predicted blocks. The biggest cause of mismatch in intra prediction is the lack of reference information. To get the predicted image using intra mode prediction, we can only 13 neighboring pixel in 4x4 luma block. It is very small amount of information compared to all pixels of the previous frame(s) in inter prediction. In generally, intra prediction coding method shows good performance for blocks with directional pattern, which is defined in intra prediction modes. But for irregular or undefined pattern, serious mismatch happens and then residual coding needs more bits. Because current intra prediction is based on only assumption of directional pattern with neighboring blocks of current block, it is impossible to predict correct image for block without directional pattern.

In this paper, the proposed intra prediction method uses the correlation between luma and chroma components. The proposed method is motivated from relation between luma and chroma component in macroblock. Why inter prediction coding method has high coding efficiency for block with irregular or unidirectional pattern is because its prediction image is not related with directional pattern. In inter coding, prediction image is obtained from reference frame and copied directly. By getting the same spatial pattern image from reference directly, the predicted image is very similar to current block. This is the most important difference between inter and intra

prediction coding. Therefore if same spatial pattern can be obtained in intra prediction, coding efficiency can be increased.

It is generally known that there is no relation between luma and chroma components. However for one macroblock of 4x4 or 16x16, spatial relation between luma and chroma exist. One macroblock size is relatively small compared to the overall image. Therefore the possibility of several objects in one macroblock is very low. That means real object may have similar luma in macroblock and chroma value and luma value can be represented by multiplication of chroma value and weight value with floating type. Therefore it is possible to predict luma block with collocated chroma value and weight value.

This paper is organized as following. Section 2 describes the proposed intra prediction coding method and section 3 shows experiment results. And section 4 leads discussion for the proposed method and finally section 5 concludes the proposed paper.

2 Proposed Methods

The main idea of the proposed method is luma blocks can be approximated as multiplication of float weight value (multiplicand) and predicted chroma blocks (multiplier). The proposed method is applied to luma component not chroma ones. Intra prediction of chroma components uses 4 prediction modes just same to the conventional H.264/AVC. In conventional intra prediction, chroma and luma components are processed separately since it is known that there is no correlation with them. In the proposed method, firstly chroma block is predicted using 4 intra prediction modes. Chroma components are divided into two sub components, U (Cb) and V (Cr). Chroma component used in prediction of luma component is made by averaging U and V components.

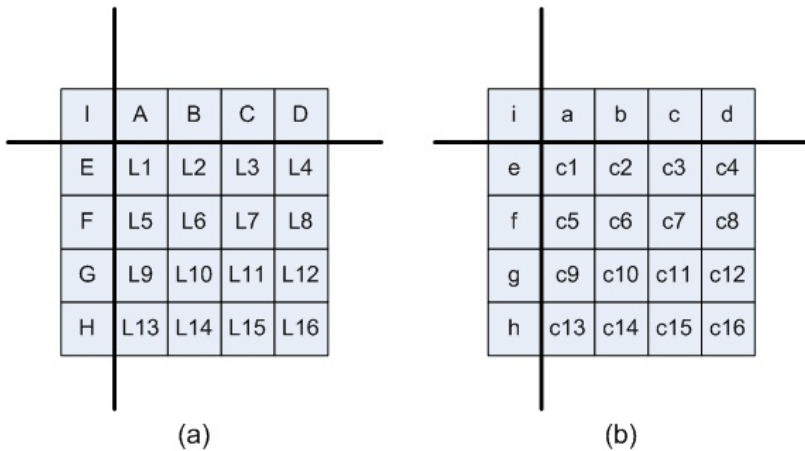


Fig. 2. (a) Original 4x4 luma block and (b) collocated upsampled chroma block

And to obtain the predicted luma blocks using the proposed method, calculation of weighted value to be used as multiplicand of the predicted chroma block is needed. Prior to weight calculation process, the predicted chroma samples and its neighboring ones are upsampled by 2 using simple bilinear interpolation because the number of chroma samples is half of luma. Fig. 2 shows 4x4 luma block and its neighboring pre-encoded luma pixels and upsampled predicted chroma block.

LN are original luma pixels and CN are upsampled collocated chroma pixels which are pre-encoded using chroma intra prediction scheme where N is 1, 2, 3,...,16. And capital alphabet characters (A, B,..., I) are pre-encoded luma neighboring pixel for current block and small alphabet characters (a, b,..., i) are pre-encoded upsampled chroma neighboring pixels. For preventing from mismatching between encoder and decoder, the proposed method also uses neighboring pixels which are predicted using intra prediction scheme. To obtain proper weight value, neighboring pixels in luma and chroma component are used. Equation 1 represents calculation of weight value used as multiplicand.

$$w = \frac{1}{9} \sum_{i=1}^9 \frac{S_i}{s_i} \quad (1)$$

where $S=\{A, B, \dots, H, I\}$, $s=\{a, b, \dots, h, i\}$.

By using neighboring pixel to obtain weight value, no additional information to be transmitted is needed. After calculation of weight value, predicted image for 4x4 luma block is obtained. Equation 2 represents 4x4 predicted luma block using weight value.

$$LN_{predicted} = w \times CN \quad (2)$$

where $LN_{predicted}$ and CN is predicted luma pixel located in N th position in 4x4 block

and CN is upsampled predicted chroma pixel located in N th position in 4x4 block

and w is weighted value calculated in Equation 1 and N is 1,2,3,...,15,16.

These methods are applied to 16x16 luma intra prediction equally, but total number of neighboring pixels to calculate weight value is 33 and maximum N is 256.

The proposed method is added to conventional intra prediction method as new additional mode and the best intra prediction mode including the proposed mode is decided using RD-constrained algorithm. To represent additional intra prediction mode, encoding scheme equal to H.264 intra prediction coding is used except total number of intra prediction mode is increased by one for both 4x4 and 16x16 luma prediction. And after obtaining predicted image, residual coding is followed to conventional H.264/AVC intra coding standard equally.

3 Experiment Results

In this paper, we add the proposed intra method into the conventional 4x4 and 16x16 luma intra prediction modes. And for evaluation of performance improvement, we test Rate-Distortion (RD) performance for 5 sequences (mobile, foreman,

football, crew, waterfall) with different size of QCIF and CIF and 4(city, crew, harbour, soccer) 4CIF sequences. The proposed method focuses on intra prediction and so we encode every frame as I-frame and don't use I_PCM option. As commented above, the proposed method is not applied into chroma component and

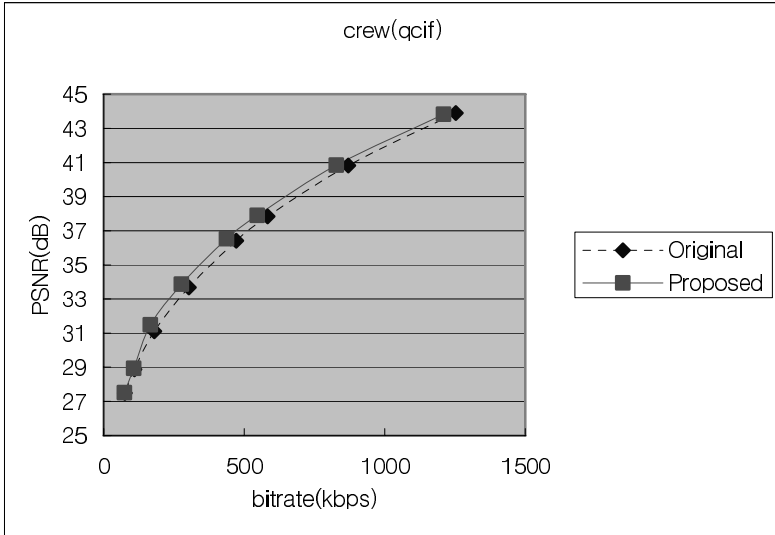


Fig. 3. RD curves for Crew (QCIF)

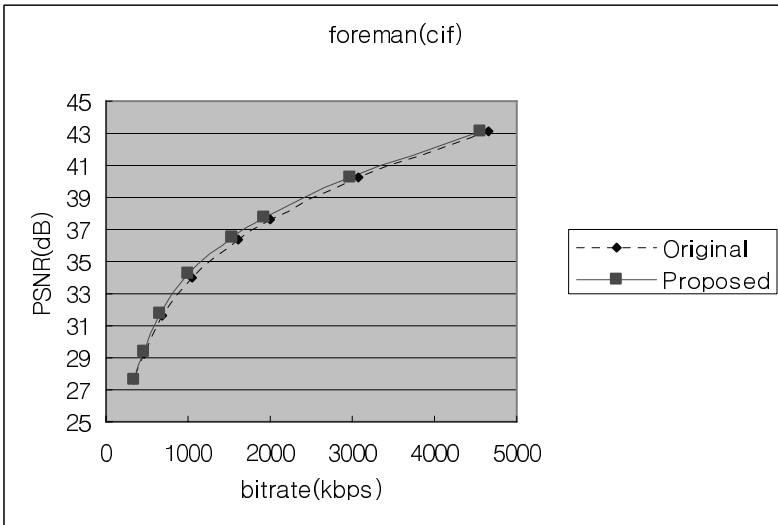


Fig. 4. RD curves for Foreman (CIF)

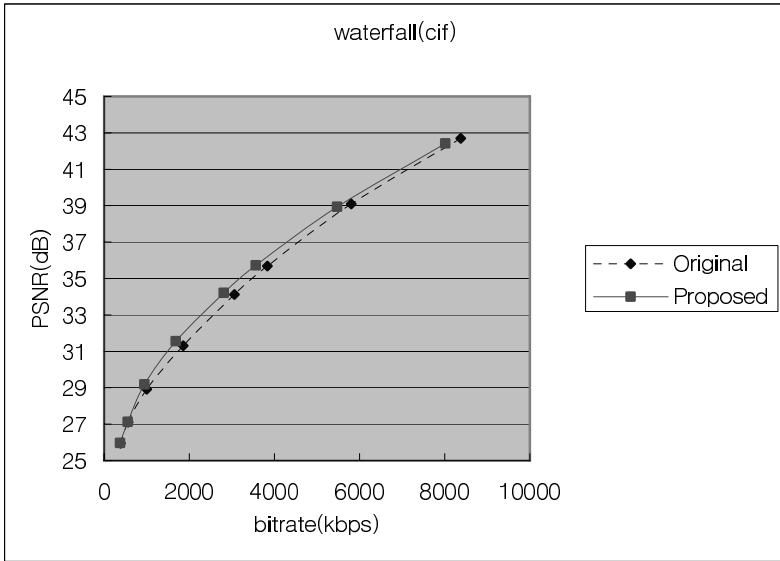


Fig. 5. RD curves for Waterfall (CIF)

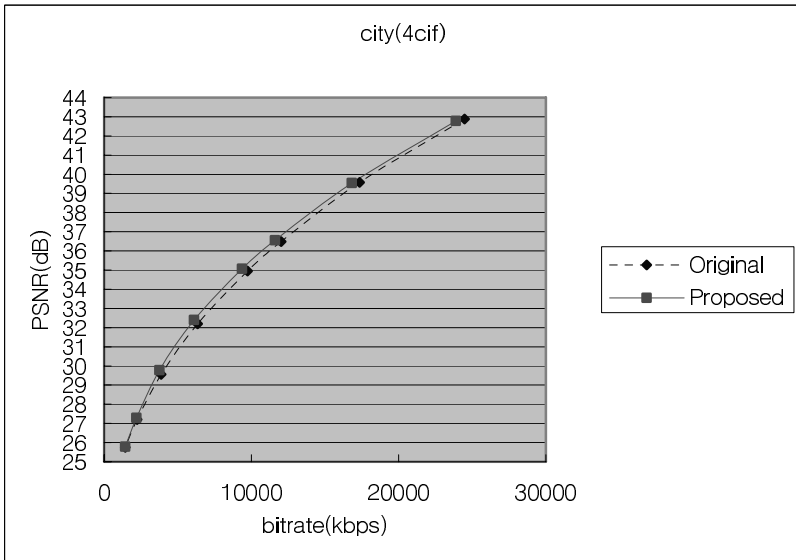


Fig. 6. RD curves for City (4CIF)

added to 4x4 and 16x16 luma intra prediction modes. We use main profile and RD optimization based on JM9.2 source. For each sequence, full number of frames is used for simulation and framerate of every sequence is 30Hz. All sequences used in experiment are I420 type which means width and height chroma components are

half of luma component. To represent coding efficiency, we use 8 QP values (20, 24, 28, 30, 34, 38, 42, 45) and measure bitrate for each QP value.

Fig. 3, Fig. 4, Fig. 5 and Fig. 6 show RD curves of conventional intra prediction method and the proposed method for Crew (QCIF), Foreman (CIF), Waterfall (CIF) and City (4CIF).

The proposed intra prediction mode shows about maximum 0.6dB enhancement of RD performance in Fig. 3 and maximum 0.5dB in Fig. 4. We identify the proposed method improves coding efficiency of intra coding by maximum 0.1~0.6dB and overall experiment result is depicted in Table 1.

Table 1. Maximum coding efficiency enhancement of the proposed method for test sequences

Resolution	QCIF					CIF					4CIF			
Name	mobile	foreman	football	crew	waterfall	mobile	foreman	football	crew	waterfall	city	crew	harbour	soccer
Max (dB)	0.5	0.6	0.5	0.6	0.6	0.5	0.4	0.5	0.5	0.5	0.4	0.3	0.4	0.3



Fig. 7. The blocks to select the proposed mode as RD-based optimal 4x4 intra prediction mode in mobile (CIF)

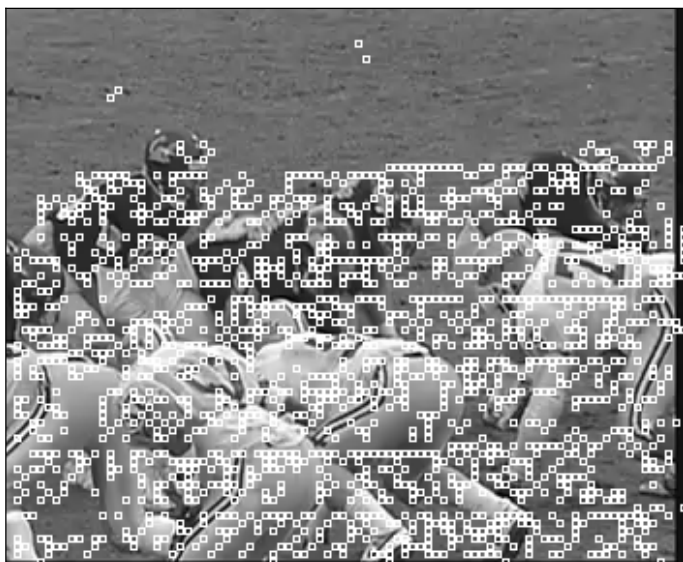


Fig. 8. The blocks to select the proposed mode as RD-based optimal 4x4 intra prediction mode in football (CIF)

In Fig. 7 and Fig. 8, they represent the blocks which choose the proposed mode as RD perspective optimal 4x4 intra prediction mode in mobile (CIF) and football (CIF) sequences. From above results the proposed mode can yield the improvement in most of intra coding blocks though its coding efficiency enhancement is even different for the block pattern.

4 Discussion

In this paper, we proposed new additional intra prediction method in H.264/AVC for more high coding efficiency of intra macroblock coding. The proposed method utilizes the correlation between luma and its collocated chroma components. In prediction of luma block, predicted luma pixel value is estimated by using multiply collocated chroma value and proper weight value and its weight value is obtained from neighboring blocks to be encoded previously. It is based on assumption that for one small block a few colors are only used in general cases. By adding the proposed method into conventional H.264/AVC intra prediction modes, overall RD performance is improved by maximum 0.6dB in sequences used for simulation.

In inter prediction coding, the predicted block is estimated by copying block which is most similar to current block in reference frame. Therefore correct prediction image is obtained if and only if similar block exists in reference frame though current block has complex or irregular pattern. The proposed method is based on prediction of similar spatial pattern, but the difference is to predict luma block from chroma in same position. For overall frame one block with 4x4 or 16x16 is not large relatively as seen

Fig. 7 and Fig. 8. Therefore in one block a few colors are used and common multiplication relation between luma and chroma value in one block exists.

Additionally in the proposed method, there is one more advantage rather than conventional intra prediction method. In conventional H.264/AVC intra prediction, the prediction image is estimated by one of prediction modes which are based on directional prediction from neighboring. This process of prediction from neighboring blocks is based on assumption that neighboring pixels have high similarity. But currently reference pixels in neighboring blocks used for directional prediction are reconstructed pixel, not original pixel. So the reconstructed pixel values are different original values and they show bigger difference between neighboring original pixel values. As bigger QP values are used, bigger difference between neighboring pixels and current original pixels is represented. And its difference value is limited by integer value level. The proposed method uses floating value as weighting value and the predicted luma value can be closer to original luma value. It can lead smaller residual values and high coding efficiency. From Fig. 3 - Fig. 6, the proposed method shows higher coding efficiency for every resolution level even if additional bits must be used for representing additional intra prediction mode. The enhancement of coding efficiency means the proposed concept about relation between luma and chroma components is very significant. In RD optimized coding routing, the proposed intra prediction method can replace conventional intra prediction modes for many blocks as seen Fig. 7 and Fig. 8. In generally more complex image is, higher coding efficiency can be obtained.

From experiment result, we identify the proposed method shows best performance middle bitrate range. The advantages of the proposed algorithm is that there is no necessary additional data bits except one bit for representing additional prediction mode for new intra prediction mode and it can improve conventional intra coding performance. In this paper, the additional computation power is a little critical problem because weight value is floating point type and for its calculation division operation is needed. Floating point type division operation can be serious from complexity point of view.

5 Conclusion

In this paper, new intra luma prediction coding method is proposed to improve coding efficiency. The proposed method utilizes relation luma and chroma in one small block and we can conclude from experimental results that the proposed relation is very significant. The main advantage of the proposed method is to improve coding efficiency of intra coding without complex modification of conventional intra coding standard. Therefore the proposed method is very useful for video coding and it can be extended to various video coding methods.

Acknowledgment

The presented research is supported by INHA UNIVERSITY Research Grant.

References

1. ISO/IEC 14496.10:2003, Coding of Audiovisual Objects. Part 10: Advanced Video Coding, 2003, also ITU-T Recommendation H.264 "Advanced video coding for generic audiovisual services."
2. ISO/IEC 15441-1:2000, Information Technology: JPEG 2000 Image Coding System. Part 1: ISO/IEC-Core Coding System.
3. ISO/IEC 13818-2:2000, Information technology -- Generic coding of moving pictures and associated audio information: Video.
4. ISO/IEC 14496-2:2004, Information technology -- Coding of audio-visual objects -- Part 2: Visual.
5. G. J. Sullivan and T. Wiegand, "Rate-Distortion Optimization for Video Compression," IEEE Signal Processing Magazine, pp. 74--90, Nov. 1998.

Fast Mode Decision for H.264/AVC Using Mode Prediction

Song-Hak Ri and Joern Ostermann

Institut fuer Informationsverarbeitung, Appelstr 9A, D-30167 Hannover, Germany
ri@tnt.uni-hannover.de
ostermann@tnt.uni-hannover.de

Abstract. In this paper, we present a new method to speed up the mode decision process using mode prediction. In general, video coding exploits spatial and temporal redundancies between video blocks, in particular temporal redundancy is a crucial key to compress video sequence with little loss of image quality. The proposed method determines the best coding mode of a given macroblock by predicting the mode and its rate-distortion (RD) cost from neighboring MBs in time and space. Compared to the H.264/AVC reference software, the simulation results show that the proposed method can save up to 53% total encoding time with up to 2.4% bit rate increase at the same PSNR.

1 Introduction

Video coding plays an important role in multimedia communications and consumer electronics applications. The H.264/AVC is the latest international video coding standard jointly developed by the ITU-T Video Coding Experts Group and the ISO/IEC Moving Picture Experts Group. It can achieve higher coding efficiency than that of previous standards, such as MPEG-4 and H.263 [1].

However, it requires a huge amount of computational loads due to use of the variable block-size motion estimation, intra prediction in P slice coding, quarter-pixel motion compensation, multiple reference frames, etc. The complexity analysis described in [1] shows that examining all possible modes takes the most time out of the total encoding time. Hence, fast mode decision making becomes more and more important.

H.264/AVC Baseline profile employs seven different block sizes for inter frames. The size of a block can be 16×16 , 16×8 , 8×16 , or 8×8 , and each 8×8 can be further broken down to sub-macroblocks of size 8×8 , 8×4 , 4×8 , or 4×4 , as shown in Fig.1. To encode a given macroblock, H.264/AVC encoder tries all possible prediction modes in the following order; SKIP, Inter 16×16 , Inter 16×8 , Inter 8×16 , Inter 8×8 , Inter 8×4 , Inter 4×8 , Inter 4×4 , Intra 4×4 , Intra 8×8 , Intra 16×16 . The SKIP mode represents the case in which the block size is 16×16 but no motion and no residual information are coded. Except for SKIP and intra modes, each inter mode decision requires a motion estimation process.

In order to achieve the highest coding efficiency, H.264/AVC uses rate distortion optimization techniques to get the best coding results in terms of maximizing

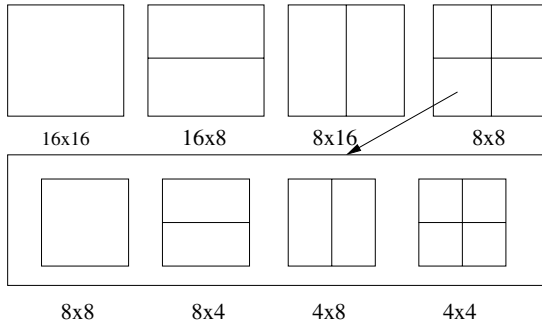


Fig. 1. Variable block sizes in H.264/AVC

coding quality and minimizing coded data bits. The mode decision is made by comparing the rate distortion cost of each possible mode and by selecting the mode with the lowest rate distortion cost as the best one.

The existing fast mode decision algorithms can be classified into two categories: The first class is to find the optimal mode by using some features, such as texture and edge information, which are computed from the raw video data. D. S. Turaga et al [5] and J. Chen et al [2] introduce the so-called mean removed mean absolute difference (mrMAD) and use the feature to make fast intra and inter mode decision. In [6], the 3×3 Sobel operator is used to get the edge map of a whole frame. The edge map and the gradient are both employed to find the best interpolation direction as the best intra mode. They also use the edge map to determine whether a macroblock is homogeneous in order to find the best inter mode. However, the algorithm has to evaluate all the pixels in the whole frame and it leads to high computational complexity.

The second class is trying to make full use of the relationship among the modes and predicts the best mode by using the already checked modes and their statistical data. A representative method [4] of such class divides all modes into 3 groups. Using one mode from each group, the best group is determined. All modes of the best group are evaluated to determine the best mode selection. Thus the number of candidate modes is greatly reduced. In [3], the most probable mode is predicted based on the observation that most modes are spatially correlated in a given frame. If the predicted mode satisfies some conditions which estimate if the predicted mode is the best mode, the encoder codes the macroblock with the predicted mode. Thus it can skip all of the calculations on other modes.

Based on the analysis above, we propose a novel algorithm to determine the best mode based on RD optimization by using the combination of spatial and temporal mode prediction. We investigate whether it is possible to temporally or spatially predict the best mode.

This paper is organized as follows: Section 2 shows a consideration about the possibility of mode prediction for fast mode decision. In Section 3, we propose a new mode decision scheme by combined mode prediction, and finally, experimental results and conclusions are presented in Section 4 and Section 5, respectively.

2 Mode Prediction

Video coding is achieved by reducing spatial and temporal redundancies between video frames. This implies indirectly that a mode of a given macroblock (MB hereafter) also might be correlated to that of MBs neighboring in space and time. It was noted that there was a spatial mode correlation between a given MB and its neighboring MBs and, therefore, it is possible to spatially predict a mode of the MB [3].

Since a video sequence contains, in general, more redundancies in time domain than in space domain, we stipulate that temporal mode correlation is higher than spatial mode correlation. Thus we consider spatial, temporal and spatial-temporal prediction of the best mode for a given MB.

In order to do that, we must answer these two questions:

1. How *high* is the correlation of spatial and temporal modes?
2. Is it necessary to consider *all* modes for the mode prediction?

Let's mark the current MB as X , collocated MB of X in the previous frame as X_{-1} and neighboring MBs as A , B , C and D (see Fig.2).

To compare correlation of both mode predictions, let's define the following 3 events:

E_S : Modes of 2, 3 or 4 MBs out of A , B , C and D are the same as the RD-optimal mode of X .

E_T : The mode of X_{-1} is the same as the RD-optimal mode of X .

E_C : $E_S \cup E_T$

Here, E_S , E_T and E_C mean spatial, temporal and combined mode events. Table 1 shows the probabilities (P_S , P_T and P_C) of each event for the video sequences *container*, *mother&daughter*, *stefan*, *mobile*, *foreman* and *coastguard*.

In Table 1, it is found that the probability of spatial mode prediction is lower than the probability of temporal mode prediction and, of course, combined mode prediction is also greater than spatial or temporal mode correlation. In the case of sequences such as *container* and *mother&daughter*, which are characterized by slow and smooth motion, the probability of a spatial mode event is similar to the temporal mode event. In the case of some sequences, such as *foreman* and *coastguard*, which are characterized by fast motion, the probability of a spatial mode event is far lower than that of the temporal mode event. The table tells us that by using combined mode correlation, the encoder can predict the best mode of a given MB more frequently than by using spatial mode correlation. From now on, the combined mode prediction will be called mode prediction.

To answer to the second question, let's calculate the probability of an event where the predicted mode of a given MB is SKIP, Inter16×16, Inter16×8, Inter8×16, Sub8×8, Intra4×4 and Intra16×16, under the condition that X has the same prediction mode with X_{-1} (see Table 2). Let's mark the

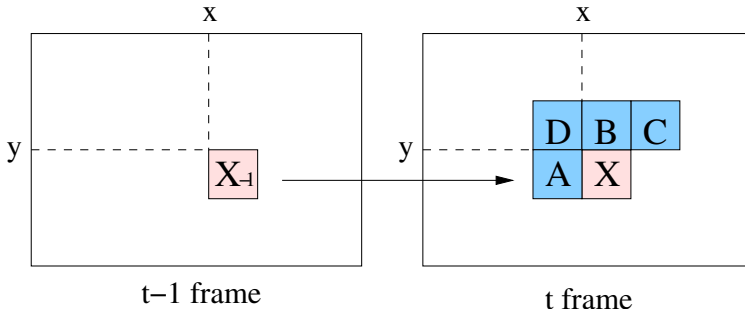


Fig. 2. The current MB X , its collocated MB X_{-1} of the previous frame, and neighboring MBs, A , B , C and D

Table 1. Comparison of an occurrence probability of spatial, temporal and combined mode events, QP (Quantization Parameter)=28, QCIF, 100 frames

Sequences	$P_S(\%)$	$P_T(\%)$	$P_C(\%)$
container	46.9	53.6	68.7
mother-daughter	33.8	40.3	56.5
stefan	13.8	31.2	41.7
mobile	10.2	27.4	35.5
foreman	8.8	29.8	34.2
coastguard	8.5	24.0	32.1

Table 2. Statistics of modewise-temporal mode correlation in case X and X_{-1} have the same RD-optimal mode (unit=%), QP=28, QCIF, 100 frames

Sequences	P_{SKIP}	$P_{Inter16 \times 16}$	$P_{Inter16 \times 8}$	$P_{Inter8 \times 16}$	$P_{Sub8 \times 8}$	$P_{Intra4 \times 4}$	$P_{Intra16 \times 16}$
container	79.5	9.4	3.6	3.7	3.7	0.0	0.1
mother-da.	64.9	16.6	6.1	6.9	5.2	0.1	0.0
stefan	21.1	32.2	10.3	15.1	20.2	0.6	0.5
mobile	17.4	27.2	13.1	10.2	31.7	0.1	0.4
foreman	12.3	39.8	13.5	11.7	22.4	0.2	0.1
coastguard	3.4	19.7	16.0	14.2	46.6	0.0	0.0

probability $P(SKIP|X=X_{-1})$ as P_{SKIP} , $P(Inter16 \times 16|X=X_{-1})$ as $P_{Inter16 \times 16}$, ..., $P(Intra4 \times 4|X=X_{-1})$ as $P_{Intra4 \times 4}$ and $P(Intra16 \times 16|X=X_{-1})$ as $P_{Intra16 \times 16}$.

As seen in Table 2, when the predicted mode of X equals the actual best mode, which can be calculated by the exhaustive mode decision of JM reference software, the occurrence probabilities of Intra4×4 and Intra16×16 are very low. This probability is also very low at other QP values, too. Therefore, we don't use the predicted modes, Intra4×4 and Intra16×16, as candidates for the best mode of a given MB, if the predicted mode is Intra4×4 or Intra16×16.

3 Fast Mode Decision by Mode Prediction

The most important thing for applying mode prediction to fast mode decision is to make sure that the predicted mode is the best mode for a given MB. So far, there have been several ways to decide whether the predicted mode is the best mode of the MB or not.

The most common method [3] adopts a threshold value derived from the RD cost which is already calculated. The threshold is set to an average of RD costs of neighboring MB with the same mode and it is compared with the RD cost of the given MB X with the predicted mode to estimate if it is the best mode. Another method [7] adopts the square of the quantization parameter (QP) as a threshold to decide whether the predicted mode is to be used.

For the sequence *foreman*, the RD cost difference between the spatially predicted mode and the optimal mode is shown in Fig 3. The size of this difference does not necessarily depend on the actual RD cost. Therefore, a threshold based on neighboring MBs or QP should not be used for evaluating the quality of the predicted mode.

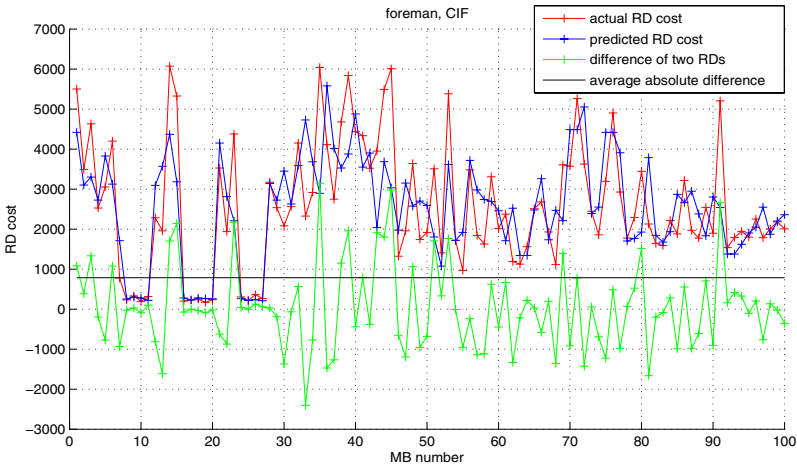


Fig. 3. Relationship between RD cost of X and average RD cost of neighboring MBs with the same mode (spatial RD cost prediction)

We use the RD cost of X_{-1} as the threshold. Fig.4 intuitively shows a relationship between the actually optimal RD cost of X and the optimal RD cost of X_{-1} when the optimal mode of X is the same as one of X_{-1} . From Fig. 5 and Table 3, it also should be noted that the correlation of both RD costs is great even in the case that the optimal mode of X_{-1} is not the same as one of X , which means that optimal RD cost of a MB can be predicted by one of the previous MB.

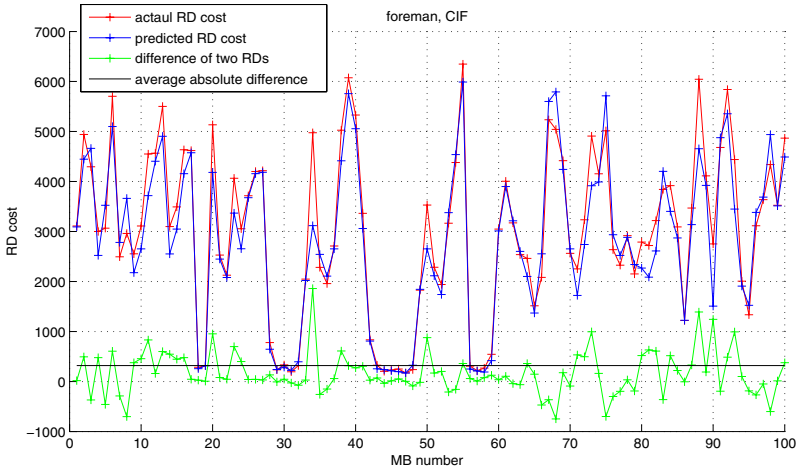


Fig. 4. Relationship between RD cost of X and RD cost of X_{-1} when the best mode of X is the same as one of X_{-1}

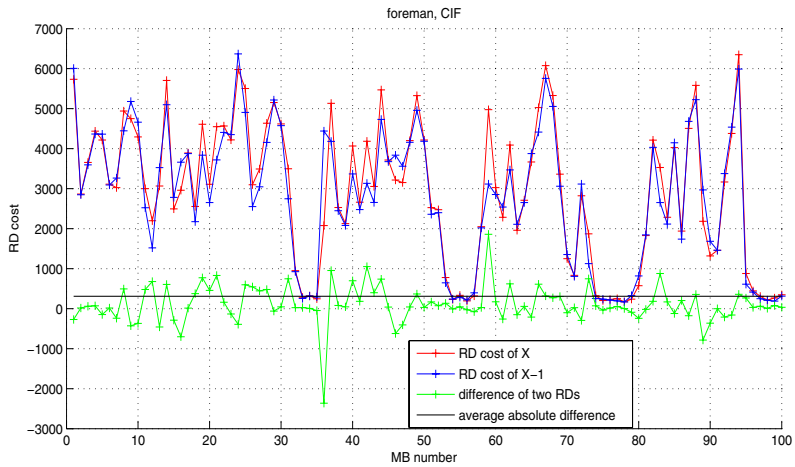
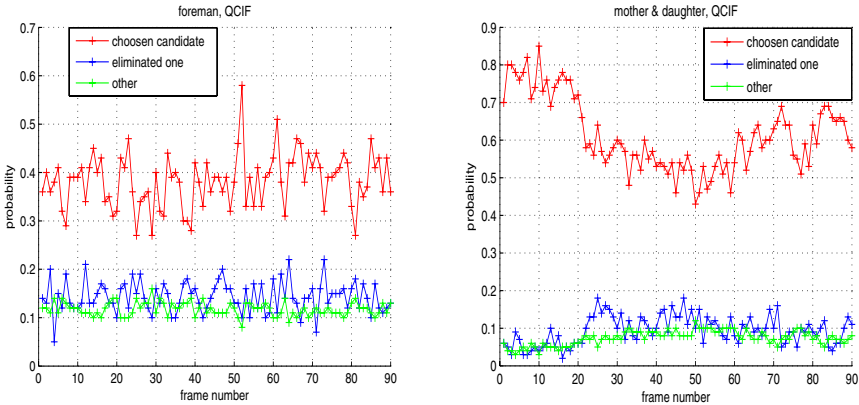


Fig. 5. Relationship between RD cost of X and RD cost of X_{-1}

Such a relationship can be seen in the comparison of the following three correlation coefficients; correlation coefficient (ρ_S) between the spatially predicted RD cost and the optimal RD cost, correlation coefficient ($\rho_{T'}$) between the actually optimal RD cost of X and the optimal RD cost of X_{-1} when the optimal mode of X is the same as one of X_{-1} , and correlation coefficient (ρ_T) between the actually optimal RD cost of X and the optimal RD cost of X_{-1} . Table 3 shows that a temporal correlation is greater than a spatial one.

Table 3. Comparison of three correlation coefficients in QCIF and CIF format

Sequences	QCIF			CIF		
	ρ_S	$\rho_{T'}$	ρ_T	ρ_S	$\rho_{T'}$	ρ_T
foreman	0.722	0.949	0.922	0.683	0.952	0.939
coastguard	0.772	0.942	0.933	0.560	0.934	0.921
stefan	0.870	0.969	0.957	0.779	0.975	0.972
mother-daughter	0.814	0.979	0.964	0.789	0.987	0.976
mobile	0.485	0.974	0.970	0.358	0.964	0.965
container	0.764	0.988	0.976	0.508	0.993	0.983
average	0.738	0.967	0.954	0.613	0.968	0.959

**Fig. 6.** Probabilities at which the chosen, the eliminated candidate or other mode is the same as the RD-optimal mode (average probability in the case of other mode)

Another problem in using mode prediction might be error propagation, due to the misprediction of the best mode. To prevent the propagation of mode prediction errors, it is expected that an exhaustive mode decision will be carried out periodically. In the experiment of the proposed algorithm, a phenomena of error propagation is likely to happen more frequently in video sequences with smooth and slow motion, resulting in some increase of the total bit rate.

The last problem of mode prediction is that using only one predicted mode to decide upon the best mode could be unstable. It has been observed that sometimes temporal mode prediction shows better result than spatial one, and also vice versa. Therefore we apply two mode candidates, m_t and m_s , predicted temporally and spatially to a given MB and choose the mode with the lower RD cost. Fig. 6 shows the three probabilities: the red curve is the probability that the chosen candidate, m_t or m_s , is the best mode, the blue curve the probability that the other mode, m_t or m_s , is the best one and the green curve shows the probability that a different mode is the best mode. As one can see, the probability that the chosen candidate is the best mode is far higher than the probability of a different mode.

The proposed algorithm is as follows:

- Step 1: if the current frame is an exhaustive mode decision frame, check all modes and stop mode decision.
- Step 2: get two predicted modes, m_t and m_s , from temporal and spatial mode predictions.
- Step 3: get the RD cost, RD_{Pred} , of the collocated MB in the previous frame with its already known best mode.
- Step 4: if both predicted modes are the same, apply it to the current MB, otherwise, compare the two RD costs by applying both and choose the better one.
- Step 5: if the chosen RD is lower than the threshold, $TH = \alpha \cdot RD_{Pred}$, set it to the best mode and stop, otherwise check all other modes (here, α is a positive constant derived from experiment).

4 Experimental Results

The proposed fast mode decision scheme was implemented in H.264/AVC reference software JM 10.1 baseline profile for performance evaluation. The experimental conditions are as follows:

Software & Profile : H.264/AVC reference software JM 10.1 Base-line
 Sequences : *container, coastguard, stefan, foreman, mobile, mother&daughter*
 Video Format : QCIF, CIF
 ME Strategy : Full Motion Estimation

The proposed algorithm was evaluated based on the exhaustive RDO mode decision of H.264/AVC in the following performance measures:

- Degradation of image quality in term of average Y-PSNR: Δ PSNR (dB)
- Increase of bit rate: +Bits (%)
- Prediction rate: PR (%)

$$PR = \frac{N_{Pred}}{N_{Total}} \times 100(\%),$$

where, N_{Total} is total number of MBs and N_{Pred} is the number of the mode prediction successes.

- Encoding time saving: TS (%)

$$TS = \frac{T_{REF} - T_{PROP}}{T_{REF}} \times 100(\%),$$

where, T_{REF} and T_{PROP} are the total encoding times of REFerence and PRO-Posed method, respectively.

In the experiment, the exhaustive mode decision is implemented at an interval of 10 frames, to prevent error propagation. α is set to 1.1.

We compared the performance of the proposed method with that of an alternative method which is based on spatial mode prediction [3]. For the QCIF video format, Table 4 shows that the proposed algorithm can achieve 44% of

Table 4. The comparison in the performance measures, QP=24, QCIF, 100 frames, where the alternative method is spatial mode prediction based method [3]

sequences	Alternative			Proposal			
	Δ PSNR(dB)	+Bits(%)	TS(%)	Δ PSNR(dB)	+Bits(%)	TS(%)	PR(%)
mobile	-0.05	3.6	24.6	0.00	1.4	42.5	45.6
stefan	-0.06	4.7	29.6	0.04	1.9	43.2	49.6
foreman	-0.01	3.5	26.5	-0.05	2.9	35.3	37.8
mother-da.	-0.03	3.9	35.3	-0.02	1.2	46.4	51.5
container	-0.02	3.7	26.2	0.01	1.7	35.1	40.7
coastguard	-0.01	2.3	23.3	-0.02	2.0	35.3	38.9
average	-0.03	3.6	27.6	-0.01	1.9	39.6	44.0

Table 5. The comparison in the performance measures, QP=24, CIF, 100 frames, where the alternative method is spatial mode prediction based method [3]

sequences	Alternative			Proposal			
	Δ PSNR(dB)	+Bits(%)	TS(%)	Δ PSNR(dB)	+Bits(%)	TS(%)	PR(%)
mobile	-0.10	3.1	29.7	-0.01	2.9	52.9	58.3
stefan	-0.08	3.7	28.6	-0.02	2.5	47.4	55.5
foreman	-0.09	3.5	38.2	-0.05	2.1	41.6	49.8
mother-da.	-0.03	3.9	46.3	-0.03	1.6	46.0	52.4
container	-0.10	3.3	37.5	-0.05	2.6	48.2	55.6
coastguard	-0.10	2.3	26.1	-0.10	1.8	51.5	63.6
average	-0.08	3.3	34.4	-0.04	2.4	47.9	55.9

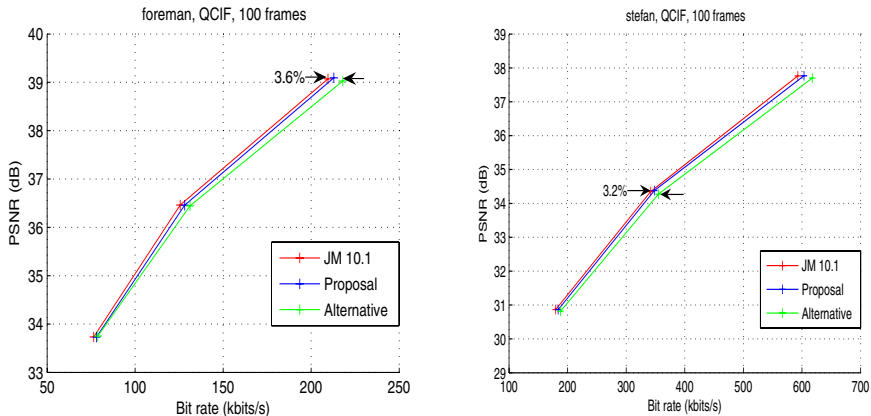


Fig. 7. Comparison of several RD (PSNR vs. bitrate) plots, QP=24, 28 and 32

average time savings in total encoding time with 0.01dB of PSNR degradation and 1.9% of extra bits. For the CIF video format, Table 5 shows better performance compared to that of the QCIF case resulting in about 48% of time saving with 0.04dB PSNR degradation and 2.4% of extra bits. The proposed algorithm shows better performance than that of the alternative algorithm [3], which is achieving about 27% of average time savings, 0.03dB of PSNR degradation and 3.6% of extra bits for QCIF sequences, and 35% of average time savings, 0.08dB of PSNR degradation and 3.3% of extra bits for CIF sequence.

Table 4 and Table 5 also show that prediction rate of the best mode depends on the contents and resolutions of video sequence, that is, how slow or fast motion is, and how fine the spatial resolution is. Fig 7 shows the rate-distortion performance of the three algorithms, the exhaustive method, alternative method (based on spatial mode prediction) and the proposed method. The curve shows that the proposed algorithm has better RD efficiency than the alternative method, achieving similar efficiency to the exhaustive method.

5 Conclusions

In this paper, we proposed a new method to speed up mode decision process using mode prediction. The proposed method determines the best coding mode of a given macroblock by predicting the mode and its RD cost from neighboring MBs in time and space. Compared to the H.264/AVC reference software, the simulation result shows that the proposed method can save up to 53% of the total encoding time with up to 2.4% bit rate increase at the same PSNR.

References

1. Ostermann, J., Bormans, J., List, P., Marpe, D., Narroschke, M., Pereira, F., Stockhammer, T., Wedi, T.: Video Coding with H.264/AVC: Tools, Performance, and Complexity, *IEEE Circuits and Systems Magazine*, Vol. 4, 1(2004) 7-28
2. Chen, J., Qu, Y., He, Y.: A Fast Mode Decision Algorithm in H.264, 2004 Picture Coding Symposium (PCS2004), San Francisco, CA, USA (2004)
3. Chang, C.-Y., Pan, C.-H., Chen, H.: Fast Mode Decision for P-Frames in H.264, 2004 Picture Coding Symposium (PCS2004), San Francisco, CA, USA (2004)
4. Yin, P., Tourapis, H.-Y.C., Tourapis, A. M., Boyce, J.: Fast mode decision and motion estimation for JVT/H.264, in *Proceedings of International Conference on Image Processing* (2003) 853-856
5. Turaga, D. S., Chen, T.: Estimation and Mode Decision for Spatially correlated Motion Sequences, *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 11, 10(2001) 1098-1107
6. Lim, K. P., Wu, S., Wu, D. J., Rahardja, S., Lin, X., Pan, F., Li, Z. G.: Fast Inter Mode Selection, *JVT-I020, JVT 9th Meeting*, San Diego, USA (2003)
7. Arsuru, E., Caccia, G., Vecchio, L. D., Lancini, R.: JVT/H.264 Rate-Distortion Optimization based on Skipping Mechanism and Clustering Mode Selection using MPEG7 Transcoding Hints, 2004 Picture Coding Symposium (PCS2004), San Francisco, CA, USA (2004)

Performing Deblocking in Video Coding Based on Spatial-Domain Motion-Compensated Temporal Filtering

Adrian Munteanu, Joeri Barbarien, Jan Cornelis, and Peter Schelkens

Vrije Universiteit Brussel,
Interdisciplinary Institute for Broadband Technology
Department of Electronics and Informatics
Pleinlaan 2, B-1050 Brussels, Belgium
{acmuntea, jbarbari, jpcornel, pschelke}@etro.vub.ac.be

Abstract. Employing block-based motion models in scalable video coding based on spatial-domain motion-compensated temporal filtering (SDMCTF) introduces artificial block-boundary discontinuities that adversely affect the video compression performance, particularly at low bit-rates. This paper focuses on the problem of deblocking in the context of SDMCTF-based video coding. One possible solution to this problem is the use of overlapped-block motion compensation (OBMC). An alternative solution is applying an adaptive deblocking filter, similar to the one used in H.264. With this respect, a novel adaptive deblocking filter, tailored to SDMCTF video coding is proposed. In terms of visual-quality, both approaches yield similar performance. However, adaptive deblocking is less complex than OBMC, as it requires up to 34% less processing time. Experimental results show that the two techniques significantly improve the subjective and objective quality of the decoded sequences, confirming the expected benefits brought by deblocking in SDMCTF-based video coding.

1 Introduction

Scalable video coding based on SDMCTF simultaneously provides resolution, quality and frame-rate scalability and yields a compression performance comparable to that of H.264, the state-of-the-art in single-layer video coding [1], [2]. These scalable video codecs typically employ block-based motion models in the motion-compensated temporal filtering process. It is well known that when such motion models fail to capture the true motion, block-boundary discontinuities are introduced in the predicted frames. These artificial discontinuities propagate in the high-pass temporal frames (H-frames) produced by temporal filtering. The H-frames are subsequently wavelet transformed and entropy coded. Due to the global nature of the spatial wavelet transform, the wavelet bases overlap the discontinuities, generating a large number of high-amplitude high-frequency coefficients, synthesizing these discontinuities. These coefficients are expensive to code, and quantizing them causes visually disturbing blocking artefacts in the decoded video, particularly when operating at low bit-rates.

One possible approach to alleviate this problem is to reduce the blocking artefacts by using overlapped-block motion compensation [3] in the predict step performed by

the temporal lifting transform. However, a known drawback of OBMC is a significant increase in the computational complexity of the employed motion model. An alternative solution is the use of an adaptive deblocking filter. This technique was adopted in H.264 to suppress the block-boundary discontinuities introduced by its block-based motion model and texture coding [4], [5]. Contrary to OBMC, which requires the execution of a large number of multiplications, H.264 deblocking can be implemented using only additions, shifts and conditional expressions. This suggests that adaptive deblocking is computationally more efficient than OBMC. However, the direct application of H.264's deblocking filter in a wavelet-based video codec is not possible, given the differences between the two coding architectures. For this reason, a novel adaptive deblocking filter, tailored to the targeted SDMCTF-based architecture is introduced in this paper. The paper compares the compression performance and the computational complexity of OBMC versus adaptive deblocking in the context of scalable SDMCTF-based video coding. The significant performance improvements offered by both approaches confirm the expected benefits brought by deblocking in wavelet-based video coding.

The paper is structured as follows: in the following section, we overview the SDMCTF video codec and describe the two approaches used to perform deblocking – OBMC and adaptive deblocking. An overview of OBMC is given in section 3, while section 4 describes the proposed deblocking filter. The experimental comparison between the two deblocking techniques is presented in section 5. Section 6 draws the conclusions of this work.

2 Performing Deblocking in SDMCTF-Based Video Coding

Motion-Compensated Temporal Filtering (MCTF) [6], [7], [8] begins with a separation of the input video sequence into even and odd temporal frames (temporal split), as illustrated in Fig. 1. The assumed motion model is a variable-size block-based motion model employing no intra-prediction. The temporal predictor performs motion-compensated prediction to match the information in frame A_{2r+1} with the information present in frame A_{2r} . The difference between the two is the high-pass temporal frame (or error-frame) H_r . Subsequently, the motion-compensated update operation inverts the prediction-error information back to frame A_{2r} , thereby producing the updated frame L_r . The update operation uses either the inverse vector set produced by the predictor, or generates a new vector set via backward motion estimation. The process iterates on the L_r frames, following the multilevel operation of the conventional lifting, thereby forming a hierarchy of temporal levels for the input video. Finally, the high-pass temporal frames H_r produced at each temporal level and the L_r frame(s) at the highest temporal level are spatially wavelet-transformed, quantized and entropy coded.

To prevent the wavelet bases from overlapping block-boundary discontinuities, a deblocking operation must precede the spatial wavelet-transform. In other words, the predicted frame needs to be free of block-boundary discontinuities prior to any spatial global transform. A possible solution is to employ overlapped-block motion-compensation in the predict step of the temporal transform. An alternative solution is to apply a deblocking technique directly after the predict-step of the temporal lifting

transform, as illustrated in Fig.1. In both approaches, the blocking artefacts are suppressed prior to wavelet-based coding of the H-frames, which is expected to result in an improved coding performance. This is to be confirmed experimentally.

Notice that one does not consider additional deblocking after the update step. Indeed, given the typically high-frequency characteristics of the H-frames, the blocking artefacts introduced in the update step are likely to be far less severe than those introduced in the predict-step.

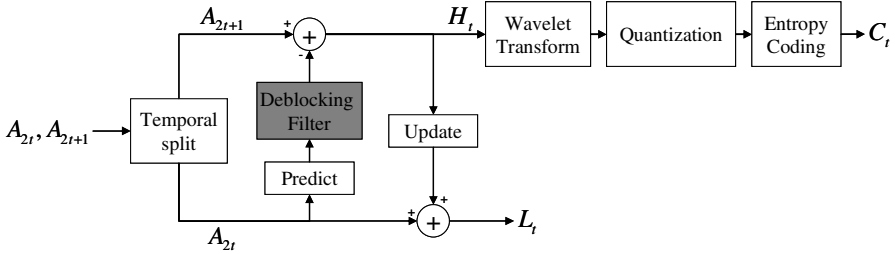


Fig. 1. SDMCTF video coding architecture; deblocking of the H_t frames can be performed in the predict step via OBMC or via adaptive deblocking preceding the spatial wavelet transform

3 Overlapped-Block Motion Compensation

The underlying assumption used in block-based motion models is that each block of pixels follows a uniform translational motion. Since this assumption is often invalid, block boundary discontinuities are created [3]. A possible solution to this problem is given by OBMC, which was first introduced in [9] and [10] and later formalized in [3].

A summary of OBMC is given next. The technique is described for a simple motion model, wherein only one block size and one reference frame are supported and a single motion vector is produced. The extension to more complex models is straightforward. The following notations are introduced:

- $I_p(x, y)$: the pixel value at position (x, y) in the predicted frame;
- $I_r(x, y)$: the pixel value at position (x, y) in the reference frame;
- B_0 : the currently visited block in the predicted frame;
- B_1, B_2, B_3, B_4 : the blocks to the left, to the right, above and below B_0 ;
- $MV_B = (mv_x^B, mv_y^B)$: the motion vector associated to block B in the predicted frame.

Using classical motion compensation, the pixel values $I_p(x', y')$ in the currently visited block B_0 are calculated as:

$$\forall (x', y') \in B_0 : I_p(x', y') = I_r(x' - mv_x^{B_0}, y' - mv_y^{B_0}) . \tag{1}$$

Each pixel belonging to B_0 is assigned the same motion vector, regardless of its position in the block. This means that the algorithm uses only a minimal fraction of the available motion information to estimate the true motion in each pixel, which is a sub-optimal approach.

OBMC on the other hand, uses the block’s own motion vector as well as the motion vectors from neighboring blocks to approximate the true motion in each pixel of the block. The pixel values $I_p(x', y')$ belonging to the currently predicted block B_0 are calculated as:

$$\forall (x', y') \in B_0 : I_p(x', y') = \sum_{i=0}^4 h_{B_i}(x', y') \cdot I_r(x' - mv_x^{B_i}, y' - mv_y^{B_i}). \tag{2}$$

The weight function $h_{B_i}(x, y)$, with $0 \leq h_{B_i}(x, y) \leq 1$ and $\sum_i h_{B_i}(x, y) = 1$, expresses the probability that MV_{B_i} corresponds to the true motion at position (x, y) . The function $h_{B_i}(x, y)$ is symmetric around the center of its associated block B_i since it is assumed that the probability that MV_{B_i} is the true motion at position (x, y) only depends on the distance of (x, y) to the center of B_i .

Equation (2) states that (a) the motion vector MV_{B_0} , associated to the current block B_0 , and the motion vectors $MV_{B_j}, 1 \leq j \leq 4$, associated to the neighboring blocks of B_0 , are all used in the motion-compensated prediction, and (b) the weight of each prediction contribution $I_r(x - mv_x^{B_i}, y - mv_y^{B_i})$ is determined by the probability that MV_{B_i} represents the true motion at position (x, y) . This approach significantly increases the accuracy of the motion-compensated prediction and prevents the manifestation of strong block-boundary discontinuities.

4 Proposed Adaptive Deblocking Filter

The adaptive deblocking filter used in H.264 [4], [5] served as basis for the design of the proposed deblocking filter. The H.264 deblocking filter consists of two parts: (1) the decision logic that determines whether filtering is turned on or off and that controls the filtering strength and (2) the filtering procedure itself. The decision logic cannot be reused in its original form given the differences between the H.264 coding architecture and SDMCTF. Indeed, in wavelet-based video coding, the block-based motion model is the only source of blocking artefacts. On the other hand, in H.264, blocking artefacts are not only caused by block-based motion compensation but also, to a larger extent, by the block-based DCT-transform and texture coding. Secondly, H.264’s motion model supports intra-prediction, while it is assumed that the one employed by the targeted wavelet-based codecs does not. Our solution therefore only reuses the filtering procedure from the original H.264 deblocking filtering, while the decision logic is adapted to the targeted SDMCTF-based video coding architecture.

The proposed deblocking filter visits the macro-blocks in raster scan order. The algorithm scans the composing sub-blocks of the currently visited macro-block and stores the coordinates of their bottom and right edges in a list of block boundaries. Block boundaries coinciding with the edges of the frame are discarded. The stored block boundaries are thereafter processed by the deblocking filter. The decision to enable or disable filtering for a block boundary $B_a B_b$ between two blocks B_a and B_b is based on the motion information associated to these blocks. Specifically, filtering is disabled only if the following conditions are met:

- B_a and B_b are predicted using the same prediction hypothesis;
- B_a and B_b are motion-compensated from the same reference frame(s);
- the difference in block motion between B_a and B_b is smaller than one luma sample.

This decision is justified since, if the above conditions hold, the blocks used to predict B_a and B_b are either adjacent or interpolated from adjacent blocks, so that block edge discontinuities do not occur.

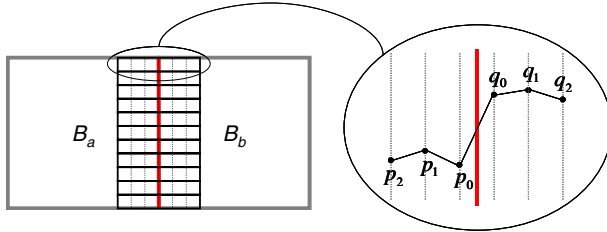


Fig. 2. A line of samples crossing the block boundary

If filtering is enabled for $B_a B_b$, the algorithm sequentially visits each line of samples crossing the block boundary, as illustrated in Fig. 2. First, the optimal filtering strength is estimated based on the local strength of the block edge discontinuity that needs to be suppressed. Let us denote by $p_2, p_1, p_0, q_0, q_1, q_2$ a line of sample values crossing the block boundary, as depicted in Fig.2. The local strength of the block edge discontinuity BDS is then estimated as:

$$BDS = |p_0 - q_0| + |p_1 - q_1|. \tag{3}$$

In our solution, the deblocking filter strength DS must be an integer value between 0 and $DS_{\max} = 36$. When DS is set to zero, filtering is turned off. The following function is used to calculate DS , given BDS :

$$DS = \min \left(DS_{\max} \cdot \frac{BDS}{\alpha}, DS_{\max} \right). \tag{4}$$

If $BDS \leq \alpha$, DS is a linear function of BDS . If BDS is larger than α , DS is set to its maximum value. The optimal value for α is determined experimentally.

Once the filtering strength is determined, the line of samples is effectively filtered. For chrominance frames, the deblocking filter only influences pixel values p_0 and q_0 . On the other hand, for luminance frames the deblocking filter can alter up to 2 pixel values on either side of the block boundary. Pixel values p_0 and q_0 are always modified by the filter, but constraint (5) below must hold before p_1 is altered and constraint (6) must be satisfied before q_1 is modified:

$$|p_2 - p_0| < \beta(DS), \tag{5}$$

$$|q_2 - q_0| < \beta(DS). \tag{6}$$

In equations (5) and (6), $\beta(DS)$ is a threshold value that depends on the deblocking strength DS . The values of $\beta(DS)$ are derived from H.264 deblocking filter's lookup table for $\beta(\text{Index}_B)$, with $\text{Index}_B = DS + 15$ [4], [5]. This explains the value of DS_{\max} . Constraints (5) and (6) ensure that q_1 and p_1 are only modified if the signal on either side of the boundary is relatively smooth. This prevents excessive blurring of highly textured regions.

Table 1. Detailed description of the filtering process for luminance frames

Modified pixel value	Conditions	Δ	c
p_0	None	$((q_0 - p_0) \ll 2) + (p_1 - q_1) + 4 \gg 3$	$c_0(DS)$
q_0	None	$-(((q_0 - p_0) \ll 2) + (p_1 - q_1) + 4) \gg 3$	$c_0(DS)$
p_1	$ p_2 - p_0 < \beta(DS)$	$(p_2 + ((p_0 + q_0 + 1) \gg 1) - (p_1 \ll 1)) \gg 1$	$c_1(DS)$
q_1	$ q_2 - q_0 < \beta(DS)$	$(q_2 + ((p_0 + q_0 + 1) \gg 1) - (q_1 \ll 1)) \gg 1$	$c_1(DS)$

Table 2. Detailed description of the filtering process for chrominance frames

Modified pixel value	Conditions	Δ	c
p_0	None	$((q_0 - p_0) \ll 2) + (p_1 - q_1) + 4 \gg 3$	$c_1(DS) + 1$
q_0	None	$-(((q_0 - p_0) \ll 2) + (p_1 - q_1) + 4) \gg 3$	$c_1(DS) + 1$

The variable-strength filtering is implemented in the same way for luminance and chrominance frames. The filtered pixel-value $p'', p'' \in \{p_1'', p_0'', q_0'', q_1''\}$ replacing the original pixel value $p, p \in \{p_1, p_0, q_0, q_1\}$ is calculated in three stages, as follows:

- In the first stage, the filtering operation is applied, leading to the filtered sample $p', p' \in \{p_1', p_0', q_0', q_1'\}$.
- From this, the difference $\Delta = p' - p$ between the filtered and original pixel values is calculated. Δ is thereafter clipped to the interval $[-c, c]$, where c is proportional to the global deblocking strength DS . This clipping stage ensures that the applied filtering strength is proportional to DS .
- The filtering process is completed by adding the clipped difference Δ' to the corresponding $p, p \in \{p_1, p_0, q_0, q_1\}$; the resulting pixel value $p'', p'' \in \{p_1'', p_0'', q_0'', q_1''\}$ replaces the original pixel value.

A detailed overview of the filtering operations is given in Table 1 for the luminance frames and in Table 2 for the chrominance frames. For each of the pixel values affected by the filtering, Table 1 and Table 2 list the conditions that need to be met

before the pixel value is effectively altered, the expression used to calculate Δ and finally, the value of c used in the clipping stage.

The threshold $c_1(DS)$ is proportional to the global deblocking strength DS . The lookup-table for $c_1(DS)$ is constructed from the lookup-table for $c_1(\text{Index}_A, B)$ [4], [5] used in the original H.264 deblocking filter, with $B=1$ and $\text{Index}_A = DS + 15$. The value of $c_0(DS)$ is initialized with the value of $c_1(DS)$ and is thereafter incremented by one for each of the conditions (5) and (6) that is satisfied.

5 Experimental Results

In a first set of experiments, we compare the compression performance obtained with a scalable SDMCTF-based video codec equipped with (a) OBMC and (b) the proposed adaptive deblocking filter against the performance obtained with the original video codec employing no deblocking. Our video codec instantiation is the SDMCTF-version of the system described in [11], which employs multi-hypothesis variable-size block-based motion estimation. The window function presented in [12] is used in the OBMC implementation. The deblocking filter's α -parameter is set to $\alpha = 28$; our experiments indicate this to be the optimum choice for a large range of sequences and bit-rates.

The comparison is performed for three CIF-resolution test sequences, "Football" (256 frames), "Foreman" (288 frames), and "Canoe" (208 frames), all at a frame rate of 30 Hz. All video codecs perform a 5-level spatial wavelet transform and a 4-level temporal decomposition. The GOP-size is fixed to 16 frames. The search ranges used in the motion estimation are $[-16, 15]$ for the first temporal level, $[-32, 31]$ for the second, $[-48, 47]$ for the third and $[-64, 63]$ for the highest level. The motion estimation is performed with quarter-pel accuracy. The results of the experiments are presented in Table 3 and Fig. 3. The reported average Peak Signal-to-Noise Ratio (PSNR) is calculated as:

$$PSNR_{avg} = (4 \cdot PSNR_Y + PSNR_U + PSNR_V) / 6 \quad (7)$$

In this equation, $PSNR_Y$, $PSNR_U$ and $PSNR_V$ are the PSNRs of the Y, U and V frame-components respectively. The results show that the codecs equipped with the proposed deblocking filtering and with OBMC yield significantly better subjective and objective quality than the original codec. The video codec equipped with OBMC outperforms the one using deblocking filtering by 0.2 dB on average. However, visual results (see Fig. 3) show that the subjective quality obtained by using the two approaches is practically the same.

In the second set of experiments, the computational complexity of OBMC and that of the proposed deblocking filter are compared. To this end, the average time needed per H-frame to perform classical motion compensation followed by the proposed

Table 3. Comparison between the original video codec and the codecs equipped with OBMC and the proposed deblocking filter

Bit-rate (kbps)	Original (dB)	OBMC (dB)	Deblocking (dB)
Football			
512	30.75	31.76	31.63
768	32.02	33.23	33.04
1024	33.38	34.61	34.38
1536	35.02	36.29	35.98
2048	36.72	37.89	37.53
Foreman			
128	31.86	32.22	32.15
256	34.55	35.00	34.87
512	37.15	37.56	37.38
768	38.61	39.24	38.98
1024	39.86	40.21	39.98
Canoa			
256	27.36	27.99	27.88
512	29.62	30.38	30.23
768	30.82	31.58	31.42
1024	31.81	32.65	32.42
1536	33.58	34.28	34.08

Table 4. Average time per H-frame needed to perform OBMC and adaptive deblocking

Sequence	MC+deblocking filtering (s)	OBMC (s)	Difference
Football	0.178	0.270	34.1%
Foreman	0.196	0.249	21.0%
Canoa	0.222	0.318	30.3%

deblocking filtering is compared to the time needed per H-frame to perform OBMC. All processing times are measured at the decoder side. A computer equipped with a Pentium-4 2.8 GHz processor and 1 GB of memory is used to run the experiments. The sequences, codec settings and target bit-rates employed in the first set of experiments are reused in these experiments. The results are shown in Table 4. The last column of the table reports the relative difference in processing time when using the adaptive deblocking filter instead of OBMC. The results show that the deblocking filter requires up to 34 % less processing time per frame than OBMC.

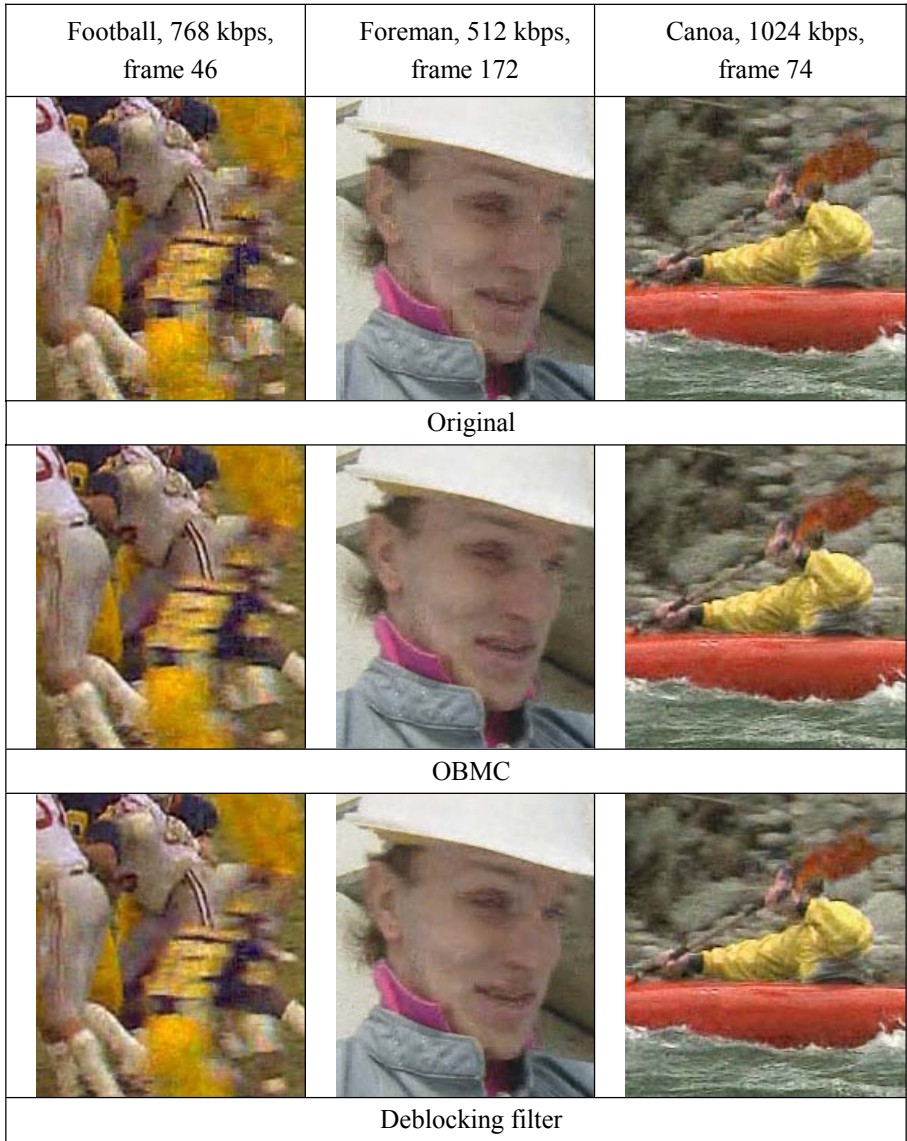


Fig. 3. Visual comparison between the original video codec and the codecs equipped with OBMC and the proposed deblocking filter

6 Conclusions

The paper focuses on the problem of deblocking in SDMCTF-based video coding employing block-based motion models and investigates two basic approaches – OBMC and adaptive deblocking – to reduce the blocking artefacts and improve the

overall coding performance of such video codecs. Inspired by the state-of-the-art H.264 deblocking filter, a novel adaptive deblocking filtering technique, tailored to scalable video coding based on SDMCTF is introduced in this paper. Experimental results show that both OBMC and the proposed deblocking filter significantly improve the subjective and objective quality of the decoded sequences. These performance improvements confirm the benefits brought by deblocking in SDMCTF-based architectures employing block-based motion models.

Compared to the deblocking filter, OBMC improves the video coding results by an additional 0.2 dB on average. However, the visual quality of the decoded sequences, obtained using both techniques is practically identical. Moreover, from a computational complexity point-of-view, the deblocking filter is clearly a better solution than OBMC, as it saves up to 34 % and on average 28% processing time. To conclude, applying the proposed deblocking filter produces the same visual quality as using OBMC but requires significantly less processing time.

Acknowledgements

This work was supported by the Flemish Institute for the Promotion of Innovation by Science and Technology (Ph.D. bursary J. Barbarien), BELSPO (IAP Phase V—Mobile Multimedia) and the Fund for Scientific Research—Flanders (FWO) (project G.0053.03.N and post-doctoral fellowships A. Munteanu and P. Schelkens).

References

1. Chen, P., Woods, J. W.: Bidirectional MC-EZBC with lifting implementations. *IEEE Transactions on Circuits and Systems for Video Technology*, 14 (2004) 1183-1194
2. Luo, L., Wu, F., Li, S., Xiong, Z., Zhuang, Z.: Advanced motion threading for 3D wavelet video coding. *Signal Processing: Image Communication, Special issue on subband/wavelet interframe video coding*, 19 (2004) 601-616
3. Orchard, M. T., Sullivan, G. J.: Overlapped Block Motion Compensation: An Estimation-Theoretic Approach. *IEEE Transactions on Image Processing*, 3 (1994) 693-699
4. Wiegand, T., Sullivan, G., Draft ITU-T recommendation and final draft international standard of joint video specification, ISO/IEC JTC1/SC29/WG11 and ITU-T SG16 Q.6
5. List, P., Joch, A., Lainema, J., Bjøntegaard, G., Karczewicz, M.: Adaptive Deblocking Filter. *IEEE Transactions on Circuits and Systems for Video Technology*, 13 (2003) 614-619
6. Ohm, J.-R.: Three-dimensional subband coding with motion compensation. *IEEE Transactions on Image Processing*, 3 (1994) 559-571
7. Pesquet-Popescu, B., Bottreau, V., Three-dimensional lifting schemes for motion compensated video compression, *Proceedings of International Conference on Acoustics Speech and Signal Processing, ICASSP 2001, Vol. 3, Salt Lake City, Utah, USA, (2001) 1793-1796*
8. Secker, A., Taubman, D., Motion-compensated highly scalable video compression using adaptive 3D wavelet transform based on lifting, *Proceedings of IEEE International Conference on Image Processing, ICIP 2001, Vol. 2, Thessaloniki, Greece, (2001) 1029-1032*
9. Watanabe, H., Singhal, S., Windowed motion compensation, *Proceedings of SPIE Conference on Visual Communication and Image Processing, Vol. 1605, (1991) 582-589*

10. Nogaki, S., Ohta, M., An overlapped block motion compensation for high quality motion picture coding, Proceedings of IEEE Symposium on Circuits and Systems, Vol., (1992) 184-187
11. Andreopoulos, I., Munteanu, A., Barbarien, J., van der Schaar, M., Cornelis, J., Schelkens, P.: In-band motion compensated temporal filtering. Signal Processing: Image Communication, 19 (2004) 653-673
12. Xiong, R., Wu, F., Li, S., Xiong, Z., Zhang, Y.-Q., Exploiting temporal correlation with adaptive block-size motion alignment for 3D wavelet coding, Proceedings of Visual Communications and Image Processing 2004, Vol. 5308, San Jose, California, (2004) 144-155

Improving DCT-Based Coders Through Block Oriented Transforms

Antoine Robert¹, Isabelle Amonou¹, and Béatrice Pesquet-Popescu²

¹ France Telecom R&D
4, rue du Clos Courtel
35512 Cesson-Sévigné Cedex
France

² TSI - ENST Paris
46, rue Barrault
75634 Paris Cedex 13
France

Abstract. This paper describes a pre-processing for DCT-based coders and more generally for block-based image or video coders taking advantage of the orientation of the blocks. Contrary to most solutions proposed so far, it is not the transform that adapts to the signal but the signal that is pre-processed to fit the transform. The blocks are oriented using circular shifts at the pixel level. Before applying these shifts, the orientation of each block is evaluated with the help of a selection based on a rate-distortion criterion. We show that the insertion of this pre-processing stage in an H.264 coder and applied on residual intra frames can improve its rate-distortion performance.

1 Introduction

Image and video compression have always aimed at finding sparse representations for the data by using transforms. The first transforms that have been introduced were separable and simple such as the DCT and the first generation wavelets. Simple means that these transforms are not optimal to represent image data in a compact way and they are often redundant, but they are fast and not too complex. This non-optimality is partly due to the fact that those transforms are not well suited to data that have discontinuities positioned along regular curves. In order to capture geometrical structures of images and video sequences, many authors proposed new transforms such as curvelets, contourlets, bandelets or directionlets, but some others improved classical transforms by taking into account the geometrical structures within data.

The curvelets which have been introduced by Candès and Donoho [1] give an optimal approximation of smooth images with C^2 edges due to their high degree of directionality. This transform requires a rotation and corresponds to a 2D frequency partitioning based on the polar coordinates, which is equivalent to a directional filter bank. It has originally been set up for the continuous case and is not easily transferable to the discrete case. In order to override this

problem, Do and Vetterli have proposed the contourlets [2] which have the same geometrical structure as the curvelets but are directly defined in the discrete case. This transform provides a multiresolution and directional analysis of a 2D signal by using a pyramidal directional filter bank. First a redundant Laplacian pyramidal multiresolution decomposition is performed, followed by a directional filter bank applied on each of the subbands. These methods involve that curvelets and contourlets are redundant and are not efficient at high bitrates.

Mallat has introduced the bandelets [3] and more recently the second generation bandelets [4], both of them being adaptive transforms. In the second generation case, a geometric orthogonal transform is applied to orthogonal wavelet coefficients by using a wavelet filter bank followed by directional orthogonal filters. Each geometric direction leads to a different transform so there is a need of edge detection and warping in order to apply the proper transform on the proper lattice.

More recently, Velisavljević and Vetterli have introduced the directionlets [5]. They work at critical sampling by applying separable filtering not only along horizontal and vertical directions but also along cosets of numerical co-lines. These numerical co-lines represent all the directions defined on an integer lattice. In order to apply the filtering along these co-lines, an integer rotation defined by the rational slope of these co-lines is performed before the horizontal filtering. But these directionlets are not suitable for block-based transforms.

Other methods use rotation of the entire image or of some parts of an image. They generally need some interpolation. Unser et al. [6] have realized fast algorithms for rotating images while preserving high quality. These rotations are decomposed in three translations performed with the help of convolution-based interpolation. The most important disadvantage of these rotation method is that, generally, information in the corners of the block is lost.

None of the methods presented before uses block-based transforms, which is the most common in the existing standards (e.g. MPEGx, H.26x and JPEG) because blocks have edges that introduce new discontinuities. Our goal is to construct a block-based rotation that keeps the shape of the block in order to apply block-based transforms.

This paper is organized as follows: in Section 2 we introduce our method and describe the way it is applied to the blocks. In Section 3, we present our selection of the orientation based on a rate-distortion criterion. Then, some numerical experiments of our pre-processing applied on residual intra frames of H.264 are presented in Section 4 before drawing the conclusions and future work in Section 5.

2 Block Oriented Transform

All the transforms presented before try to adapt to the signal but each of them needs some floating-point processing. Contourlets and curvelets are redundant, bandelets need an edge detection and warp the data, rotations need interpolation. All these drawbacks are in contrast with a perfect reconstruction. The method we propose is a pre-processing of images or video sequences that takes advantage of the geometrical structure of the data without warping or interpolation.

The best-known block-based transform is the DCT which is used in most of the image or video standards, like JPEG [7], MPEGx, H.26x like H.264-MPEG4/AVC [8]. In this transform, the basic element is a block of coefficients which can be of size 8x8 for the floating-point DCT like in JPEG or 4x4 for the integer DCT in H.264. These blocks can be of different nature: real data coming from the real images (JPEG) or residual data coming from a prediction (intra predicted blocks of H.264). In each case these blocks still show some regular pattern like presented on Figure Fig.1.



Fig. 1. Part of Flower (CIF) and its residual after intra prediction in H.264

Our pre-processing focuses on exploiting the orientation of these blocks without using any rotation scheme which implies interpolation, but it is done by some circular shifts at the pixel level.

2.1 Considered Orientations

We first define all the orientations that are going to be used in our scheme, for a given block size. The orientations are selected in tables defining intervals. Each of the intervals corresponds to a range of pre-calculated angles for the block. In the 4x4 case, we define seven states : three states for the positive angles (cf Fig.2), three states for their opposites and one more state referring to non-oriented blocks and blocks which have not their directions in the defined intervals (cf Tab.1).

2.2 Block Orientation

After an orientation has been defined for the block, using this table, we have to pre-process it according to the corresponding orientation.

In state 0 nothing is done, because either the blocks are non-oriented (if their direction is 0° or 90°), or the blocks do not have a direction that is included in an interval defined before.

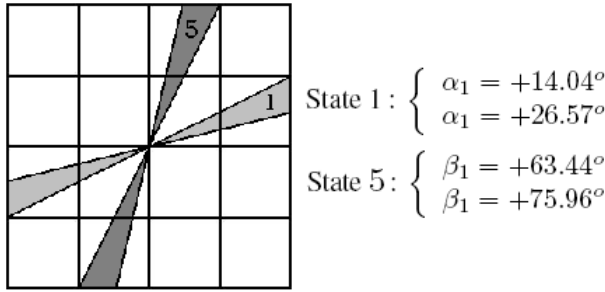


Fig. 2. Two states for the positive angles in the 4x4 case

Table 1. All states in the 4x4 case

State	Interval	State	Interval
1	[+14°; +27°]	2	[-27°; -14°]
3	[+40°; +50°]	4	[-50°; -40°]
5	[+63°; +76°]	6	[-76°; -63°]
0			all other interval

In all the others states, some circular shifts at the pixel level are applied in order to perform a reorientation of the block. These circular shifts enable us to override the problem of interpolation that exists in a rotation scheme. Moreover, by this simple pixel rearrangement we simulate the corresponding rotation without creating “holes” in the corners of the blocks. It is important to note that our scheme just simulate a rotation by using shearing, it is not a real rotation (which means to combine several shearing).

In state 1 (cf Fig.3) a circular shift on the first two pixels of the first line of the block is performed and in state 2 (its opposite) on the last two pixels of the first line. In states 5 and 6 the same rearrangements are employed but on the first two columns (cf Fig.3). States 3 and 4 use more complex pixel rearrangements. State 3 corresponds to circular shifts applied on the first pixel of the first line and on the last pixel of the last line before applying the same circular shift as in state 1. State 4 is the same as state 3 but applied to the columns: shift first and last pixels of the first and last columns before applying state 2 (cf Fig.3).

Figure Fig.3 shows that the direction of the block is coming back to 0° or 90° after the rearrangement had been applied. The circular shifts have carried out the simulation of the “rotation” without its disadvantages.

All the defined states can be summarized by two circular shifts: the first one is applied on the first (1,5) or the last two pixels (2,6) of the first line (1,2) or column (5,6), the second one is applied on the first pixel of the first line (3) or column (4) and on the last pixel of the last line (3) or column (4) (cf Tab.2).

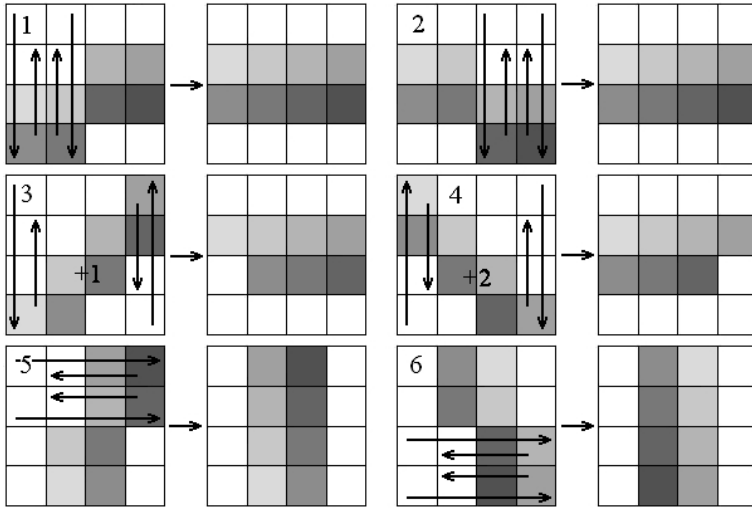


Fig. 3. Circular shifts in the 4x4 case

Table 2. All circular shifts in the 4x4 case

State	Nb pixels		Line/Column	
	First	Last	First	Last
1/5	2		X	
2/6		2		X
3/4	1	1	X	X
	Same as State 1/2			

3 Orientation Selection

Before we apply the pseudo-rotation, we have to select the proper orientation for each block of our images or video sequences.

In order to select this orientation for each block, we have developed a method based on the rate-distortion optimization (RDO) of H.264 [9]. This optimization consists in testing all the combinations of available modes and coding the macroblocks with the one that gives the best performance: the lowest distortion for a given bitrate or the lowest bitrate for a given distortion.

The coding cost of a macroblock thus depends on two variables: rate and distortion. The macroblock rate is, in all the cases, the sum of the blocks rates that compose it. And the distortion is always the total distortion of the macroblock whatever the blocks sizes, it is given by the quadratic error of the reconstructed macroblock:

$$D = \sum_{m=0}^{15} \sum_{n=0}^{15} (i_{MB}(m, n) - \hat{i}_{MB}(m, n))^2 \tag{1}$$

where $i_{MB}(m, n)$ is the pixel (m, n) of the original macroblock and $\hat{i}_{MB}(m, n)$ the corresponding in the reconstructed macroblock.

In H.264 and in intra coding, each 4x4 block is tested with all the nine prediction modes [8] before being coded. The best prediction mode in the rate-distortion sense is kept for the real coding stage. Compared to a traditional coder, our pre-processing just adds combinations to be tested.

All of the orientations defined in section 2.1 are tested in all the H.264 existing prediction modes. In other words, instead of testing each H.264 prediction mode once, they are tested seven times, corresponding to the seven possible orientation states (cf Tab.1). The best prediction mode with the best orientation in the rate-distortion sense is kept for coding.

It is interesting to note that in state 0, none of the orientations being selected, H.264 is applied as such.

This selecting method is efficient but complex in terms of rate-distortion evaluations: for example, for an intra block, as 9 prediction modes are defined in H.264, a total number of 63 combinations of prediction modes with orientations have to be tested per block. However this increase of complexity is only applied on rate-distortion selection of the intra prediction modes of H.264 that represent only a small part of the coding stage of H.264. The total complexity of the H.264 algorithm is not much increased.

4 Experimental Results

We have evaluated our pre-processing with the selecting method described above: rate-distortion based selection. In these tests, we do not encode yet the orientation states (estimated at 1-2% of the bitstream, like the intra prediction modes in H.264).

All the experiments have been conducted inside an H.264 video coder [8] the JM10 [10] provided by the JVT, in Main profile at level 4.0, but only on residual intra frames. In order to have only 4x4 blocks, we have forced the I4MB mode which corresponds to a 4x4 intra-prediction with the 4x4 integer DCT. The sequences are generated by varying the QP for intra slices over all available values (0-51 and fixed at 28 for inter slices). They are then made up of three images: an I, a P and a B.

The results for the sequence Container in CIF format at 15Hz are shown on Figure Fig.4 and those for the sequence Mobile&Calendar in CIF at 15Hz on the Figure Fig. 5.

One can see on these figures that our pre-processing improves H.264 coding at all bitrates. The PSNR improvement over H.264 ranges from 0.16dB for the sequence Container and from 0.38dB for the sequence Mobile&Calendar at 150kbits/s, to up than 1dB at high bitrate in both case (after 2500-3000kbits/s).

Similar results have been obtained with a large number of sequences like Akiyo, Foreman, Bus, Tempete as shown in Table Tab.3. For example, the sequence Tempete generates a gain of 0.30dB compared to H.264 at 200kbits/s, and a gain higher than 0.5dB beyond 3100kbits/s or for a QP lower than 9.

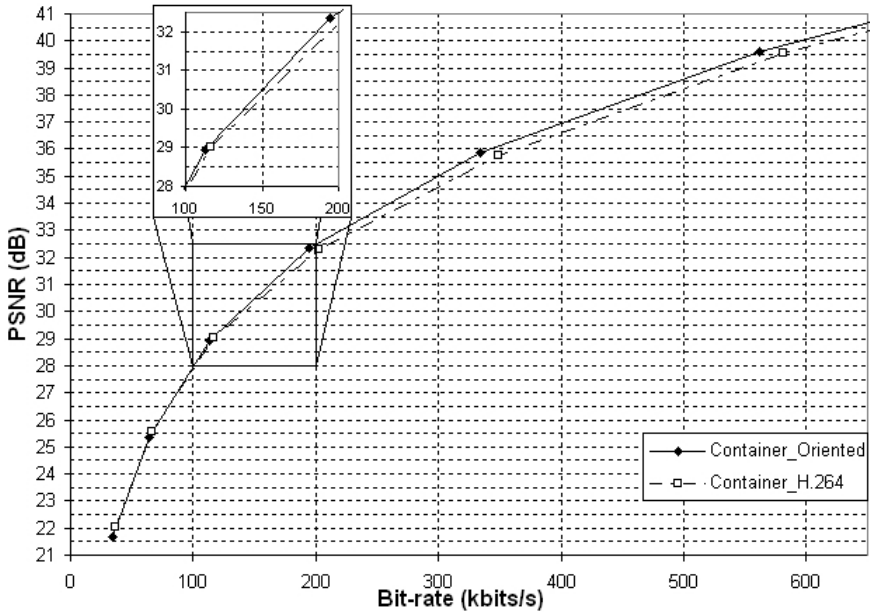


Fig. 4. Results for the sequence Container (CIF)

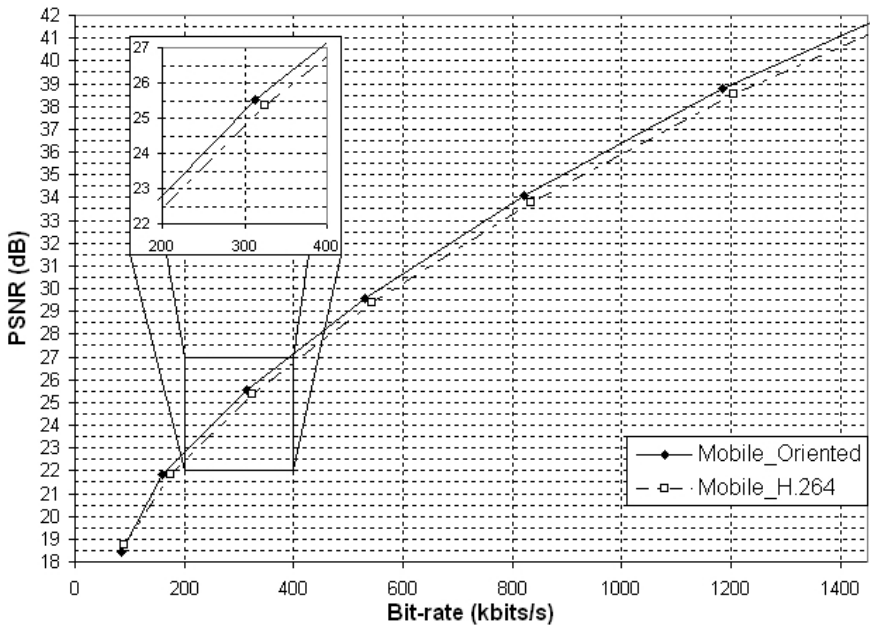


Fig. 5. Results for the sequence Mobile&Calendar (CIF)

Table 3. Results for other sequences

Sequence	$d = 200\text{kbits/s}$	$\Delta_{PSNR} > +0.5\text{dB}$	
CIF	$\Delta_{PSNR} = (\text{dB})$	$d > (\text{kbits/s})$	$QP <$
Akiyo	+0.10	1300	10
Bus	+0.19	2600	9
Flower	+0.32	3000	10
Football	+0.24	2600	6
Foreman	+0.23	1800	9
Tempete	+0.30	3100	9
Sequence	$d = 80\text{kbits/s}$	$\Delta_{PSNR} > +0.5\text{dB}$	
QCIF	$\Delta_{PSNR} = (\text{dB})$	$d > (\text{kbits/s})$	$QP <$
Carphone	+0.13	500	9
Foreman	+0.23	600	7

5 Conclusion and Future Work

We have introduced a pre-processing tool for image and video block-based coders, based on block reorientations without any use of segmentation, warping or interpolation. First, an orientation is selected for each 4x4 block of our images or video sequences by using a rate-distortion optimization. Then the block is straightened out according to this orientation. This pseudo-rotation (shearing) is obtained by applying simple circular shifts on the pixels of the blocks. We have shown that this pre-processing can improve significantly DCT-based coders like H.264.

Our next step will be to effectively encode the orientation states for each block: we plan to use CABAC [11] and a method very similar to the one used in H.264 for the encoding of intra prediction modes. We will also work in reducing the complexity of the algorithm by using some measure of the orientation. We are also going to introduce the 8x8 (FRExt [12] only) and 16x16 cases and to extend our pre-processing to chroma components and inter frames.

References

1. Candès, E., Donoho, D.: Curvelets, multiresolution representation, and scaling laws. In: Wavelet Applications in Signal and Image Processing VIII. Proc. SPIE 4119 (2000)
2. Do, M., Vetterli, M.: The contourlet transform : An efficient directional multiresolution image representation. IEEE Transactions on Image Processing **14** (2005) 2091–2106
3. Le Pennec, E., Mallat, S.: Sparse geometric image representations with bandelets. IEEE Transactions on Image Processing **14** (2005) 423–438
4. Peyré, G., Mallat, S.: Discrete bandelets with geometric orthogonal filters. IEEE International Conference on Image Processing (ICIP'05) **1** (2005) 65–68
5. Velisavljević, V., Beyerull-Lozano, B., Vetterli, M., Dragotti, P.: Directionlets : Anisotropic multi-directional representation with separable filtering. IEEE Transactions on Image Processing (2005)

6. Unser, M., Thévenaz, P., Yaroslavsky, L.: Convolution-based interpolation for fast, high-quality rotation of images. *IEEE Transactions on Image Processing* **4** (1995) 1371–1381
7. ISO/IEC 10918: Digital Compression and Coding of Continuous-Tone Still Images, ISO/IEC 10918, JPEG (1994)
8. JVT - ISO/IEC 14496-10 AVC - ITU-T Recommendation H.264: Advanced video coding for generic audio-visual services, JVT - ISO/IEC 14496-10 AVC - ITU-T Recommendation H.264, Draft ITU-T Recommendation and Final Draft International Standard, JVT-G050r1 (2003)
9. Wiegand, T., Schwarz, H., Joch, A., Kossentini, F., Sullivan, G.: Rate-constrained coder control and comparison of video coding standards. *IEEE Transactions on Circuits and Systems for Video Technology* **13** (2003) 688–703
10. : Joint model 10 (2006) <http://iphome.hhi.de/suehring/tml/index.htm>.
11. Marpe, D., Schwarz, H., Wiegand, T.: Context-based adaptive binary arithmetic coding in the h.264/avc video compression standard. *IEEE Transactions on Circuits and Systems for Video Technology* **13** (2003) 620–636
12. JVT - ISO/IEC 14496-10 AVC - ITU-T Recommendation H.264 Amendment 1: Advanced Video Coding Amendment 1 : Fidelity Range Extensions, JVT - ISO/IEC 14496-10 AVC - ITU-T Recommendation H.264 Amendment 1, Draft Text of H.264/AVC Fidelity Range Extensions Amendment (2004)

Improvement of Conventional Deinterlacing Methods with Extrema Detection and Interpolation

Jérôme Roussel^{1,2}, Pascal Bertolino², and Marina Nicolas¹

¹ ST Microelectronics S.A., 12 Rue Jules Horowitz B.P. 217, GRENOBLE - France

² Laboratory of Images and Signals, INPG, BP 46-38402 St Martin d'Hères - France

Abstract. This article presents a new algorithm for spatial deinterlacing that could easily be integrated in a more complete deinterlacing system, typically a spatio-temporal motion adaptive one. The spatial interpolation part often fails to reconstruct close to horizontal lines with a proper continuity, leading to highly visible artifacts. Our system preserves the structure continuity taking into account that the mis-interpolated points usually correspond to local value extrema. The processing is based on chained lists and connected graph construction. The new interpolation method is restricted to such structures, for the rest of the image, a proper traditional directional spatial interpolation gives satisfactory results already. Although the number of pixels affected by the extrema interpolation is relatively small, the overall image quality is subjectively well improved. Moreover, our solution allows to gain back one of the major advantages of motion compensation methods, without having to afford their complexity cost.

1 Introduction

The video signal is transmitted over the world in interlaced frames. For technical reasons, the TV signal frame frequency was selected according to the frequency of the electrical power supply and the requirements concerning large area flicker. Interlacing made it possible to cope both with the frame rate and the resolution requirements. However, new flat panels like plasma or L.C.D are progressive ones and thus require the display of the whole image at time t . Moreover, interlacing also causes flicker on objects containing high horizontal frequencies. There is thus a high interest in deinterlacing methods that allow a conversion from interlaced to progressive. They can be classified in two major families, the methods without motion compensation and the ones with motion compensation [1]. We will focus here on the methods without motion compensation. Those methods can in their turn be split in temporal, spatial and spatio-temporal adaptive methods. The adaptive method consists in going towards the temporal method in areas where there is no movement and towards the spatial method in moving areas [2] - [3]. The spatial interpolation part has its own limitations, mainly on the rendering of close to horizontal lines and the technique proposed in this article addresses this particular point. First, the existing methods are reviewed showing that the limitation mentioned cannot be easily overcome. Then, in a second part we describe

our method, based on a new extrema detection and interpolation principle. The major improvements reached are then showed. Finally, we conclude on the cost of the method and give some hints about the remaining work.

2 Existing Methods

From now on in the article, f_{in} represents the interlaced input image and \tilde{f} the interpolated output image. (i, j) are the spatial coordinates where i represents the line position number and j the column position number. f_{in} is defined only for half of the lines, i.e. for i even or odd. Many methods and solutions have been proposed to perform spatial interpolation [4]. The first very basic one consists in using the average of the pixels above and below the missing one to interpolate the missing pixel, i.e. :

$$\tilde{f}(i, j) = \frac{(f_{in}(i - 1, j) + f_{in}(i + 1, j))}{2}. \quad (1)$$

This method does not make it possible to reconstruct high frequencies (contours) in a sharp way. Typically, it can introduce staircase contours, flicker or blur. To improve this algorithm, the next idea was to make the interpolation along the direction of contours, using the so-called E.L.A method (Edge-based Line Averaging) [5]. This latter method detects the best direction Dir for interpolation within a window centered on the missing pixel and then makes the interpolation according to the found direction:

$$\tilde{f}(i, j) = \frac{f_{in}(i - 1, j - Dir) + f_{in}(i + 1, j + Dir)}{2}. \quad (2)$$

Although it leads to a better interpolation of contours, the method still has several limitations. Indeed, the correlation is done at the local level and remains quite sensitive to noise. The direction of contours thus happens to be wrong which can lead to very annoying artifacts since it disrupts the structure of thin lines or contours. Many alternatives of this method [6] make it possible to correct wrong direction interpolation for a majority of the pixels of the image, for instance by computing the correlation between groups of pixels instead of pixel to pixel (*figure 1*). However, the results of these methods always remain dependent on the size of the window used, that determines the maximum angle allowed for the reconstruction of the contours. On the other side, the larger the window is, the higher the risk of bad interpolation [7]. Different existing features and metrics try to control an adequate window size and introduce weights to reduce the number of false directions [8] - [9]. But methods used to calculate this weights significantly increase the complexity of the solution. Still, all these alternatives only bring a final minor improvement and do not allow to reconstruct correctly close to horizontal lines and structures. This point is even more annoying in real time where moving horizontal lines do not only look disrupted, but also instable and highly flickering.

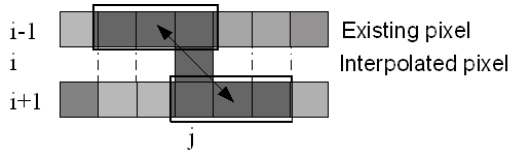


Fig. 1. Principle of modified E.L.A.

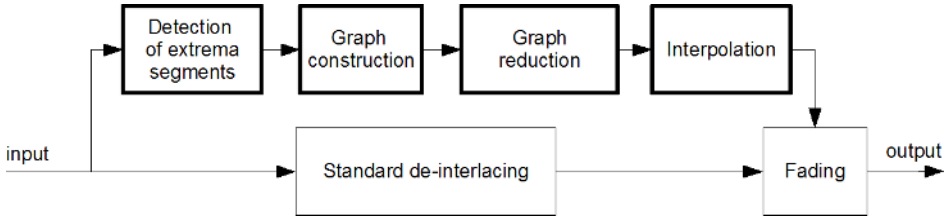


Fig. 2. Block diagram of the process. The stages of the proposed method are in bold.

Our goal was thus to propose a system that solves this problem significantly without having to pay for a motion compensation solution. The method proposed here overcomes the limitation of the current directional spatial interpolation method since it is not based anymore on a searching window. It also only affects pixels values in image areas with given characteristics that cannot be handled properly in existing methods. Typically it overrules the wrong interpolation of close to horizontal structures, keeping the results of the traditional spatial interpolation where the results are already satisfactory. It can be added to a classical spatial deinterlacing method that can itself be integrated in a motion adaptive one (*figure 2*).

3 Extrema Detection

The existing methods are not able to respect the continuity of close to horizontal thin structures. It is all the more awkward as the visual artifacts due to this problem are often very visible (disconnection, erroneous interpolation), as shown in the figure 3.c.

By comparing the modulus of the Fourier transform of a frame and the one of a whole image, one can observe in the case of the frame, on the one hand the spectrum folding phenomenon constituting the aliasing, and on the other hand the loss of the horizontal high frequencies. The difficulty consists in locating and reconstructing the continuity of these high frequencies structures which were partially destroyed and systematically disconnected by horizontal under-sampling (*figure 3.b*). Those correspond to local minima or maxima of the intensity function, in the vertical direction. The detection of these local extrema is carried out

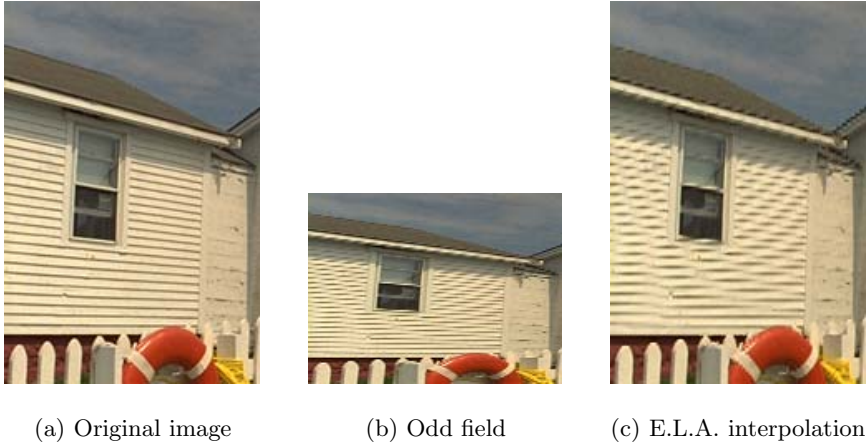


Fig. 3. The algorithms which use a search window cannot accurately rebuild the extrema-type structures

on the known lines of the image by comparing the value of each pixel $f_{in}(i, j)$ with the values of the closest lower and higher lines $f_{in}(i-2, j)$ and $f_{in}(i+2, j)$. Let \mathcal{H} be the set of all maxima pixels and \mathcal{L} the set of all minima pixels:

$$\mathcal{H} = \{f_{in}(i, j) / f_{in}(i, j) > \max(f_{in}(i-2, j), f_{in}(i+2, j)) + T\} \quad (3)$$

$$\mathcal{L} = \{f_{in}(i, j) / f_{in}(i, j) < \min(f_{in}(i-2, j), f_{in}(i+2, j)) - T\} \quad (4)$$

T is a threshold value of minimum contrast ($T = 16$ in our experiments).

4 Segments and Associated Data Structure

On the same line, the extrema of the same type can form connected components (called segments) according to the horizontal 2-connectivity (*figure 4*). As the continuation of the method is not founded on the traversing and the processing of pixels but on the traversing and the processing of segments, the traditional two-dimensional image structure is not suitable any more and is replaced with a high level structure: namely the segment. Each segment is an entity characterized



Fig. 4. Example of segments of the same type extracted (in black). The gray lines are known, the white lines are to be interpolated.

by its coordinates (line, starting column), its length and type (minimum or maximum). The chosen data structure is a compromise between the memory size needed and the complexity to traverse \mathcal{H} and \mathcal{L} . The adopted solution is an array of lines (an entry for each frame line), each line being a chained list of the segments extracted in the corresponding frame line.

5 Construction of Connected Graphs

This stage has two goals : first of all, a practical goal to fill the structure presented above with all the segments, so that it can be easily traversed. Then a functional goal to interconnect the segments of \mathcal{H} (resp. of \mathcal{L}) to constitute one or more connected graphs of maxima (resp. one or more connected graphs of minima). The traversing of the elements of $(\mathcal{H} \cup \mathcal{L})$ in the array of chained lists is done in the video scanning direction.

A segment S on a line i has at most 6 direct neighbors of the same type: 3 on the west side and 3 on the east side, 2 on each line $i - 2, i$ and $i + 2$ (figure 5). The distance between two neighboring segments of the same type S_1 and S_2 is the Euclidean distance $d(S_1, S_2)$, calculated between the closest extremities of S_1 and S_2 .

For a given side (west or east), S is connected to its closest neighbor. If two neighbors are closest at the same distance, S is connected to both (figure 6). Connections are bidirectional. An adaptive threshold is used in order to avoid not very reliable (too long) connections. L_1 and L_2 being the respective lengths of S_1 and S_2 , the distance $d(S_1, S_2)$ must check the following condition so that S_1 and S_2 are connected:

$$d(S_1, S_2) < \min(L_1, L_2) + \delta \tag{5}$$

In our experiments δ is fixed at 2. Connections of a segment with its neighbors are stored in the segment itself (as pointers to segments). It should be noted that



Fig. 5. The white segment has 5 direct neighbors



Fig. 6. Connected graph derived from the segments of figure 4

because of the intrinsic nature of local extrema, a segment cannot have more than two neighbors for a given side.

6 Graph Reduction

The whole of the extracted graphs cannot be interpolated as it is. Certain connections must be removed (*figure 7*), which causes the division of a graph into several subgraphs. The graphs comprising only one segment or which are on only one line are also removed. The suppression of connections must on the one hand privilege subgraphs having each one a prevalent direction and on the other hand remove the false positive ones (connections performed wrongly).

Let NW, W, SW, NE, E, SE be the directions associated to the 6 possible connections for a segment. When performing the in-depth traversal of the graph, let us call input direction the one by which the segment is reached. This direction is non-existent for the starting segment of the traversal. The output directions correspond to all connections of the segment except the input direction. The simplification complies with the following rules :

1. If there are 2 output connections on the same side, they are removed. This rule makes it possible not to connect potentially different structures wrongly.
2. If an output connection is on the same side that the input connection, it is removed. This rule makes it possible to preserve only structures that are stretched and not zigzag-like.

The subgraphs resulting from the traversal and the reduction rules are trees with only one branch. Long rectilinear or curve structures can be thus reconstituted.



Fig. 7. The remaining connections after the reduction of the graph of figure 6

7 Interpolation

The interpolation is the last stage (*figure 8*). It is carried out using a forward (from west to east) traversing of each branch.

Let S_1 and S_2 be two connected segments described by their lengths L_1 and L_2 , their starting coordinates (Y_{S_1}, X_{start_1}) and (Y_{S_2}, X_{start_2}) . The pixels to be interpolated with our method correspond to the segment S_I whose extremities



Fig. 8. The segments in dark gray represent the pixels interpolated thanks to the black neighboring segments

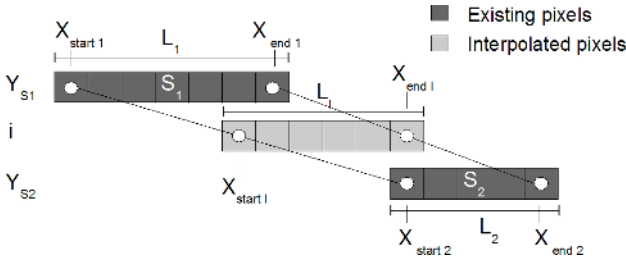


Fig. 9. Interpolation principle

X_{start_I} and X_{end_I} are linearly interpolated from the extremities of the segments S_1 and S_2 (figure 9).

$$X_{start_I} = X_{start_1} + \left\lfloor \frac{X_{start_2} - X_{start_1}}{2} \right\rfloor \quad (6)$$

$$X_{end_I} = X_{end_1} + \left\lfloor \frac{X_{end_2} - X_{end_1}}{2} \right\rfloor \quad (7)$$

$$\begin{aligned} \tilde{f}(i, j) &= \frac{1}{2} f_{in} \left(Y_{S_1}, X_{start_1} + E \left(\frac{j \times L_1}{L_I} \right) \right) \\ &+ \frac{1}{2} f_{in} \left(Y_{S_2}, X_{start_2} + E \left(\frac{j \times L_2}{L_I} \right) \right) \quad (8) \\ &j \in [X_{start_I}, X_{end_I}] \end{aligned}$$

The function E returns the nearest integer of its argument. L_I is the size of the segment to interpolate $X_{end_I} - X_{start_I} + 1$. Y_{S_1} and Y_{S_2} represent the ordinates $i - 1$ and $i + 1$ (or $i + 1$ and $i - 1$).

8 Results

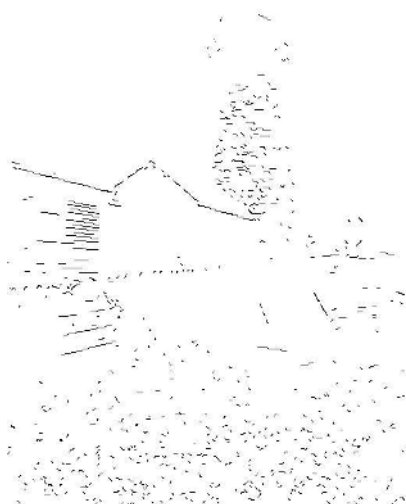
The tests were carried out on a rather broad set of sequences, some coming from originally interlace material, others from originally progressive material that has been re-interlaced. The latter ones can be used as reference pictures to estimate the quality of the interpolation. The extrema pixels were interpolated following



(a) Original image



(b) De-interlacing with our method

(c) Connection of minima
in the full size image(d) Connection of maxima
in the full size image**Fig. 10.** Reconstruction of thin structures using half of an original image

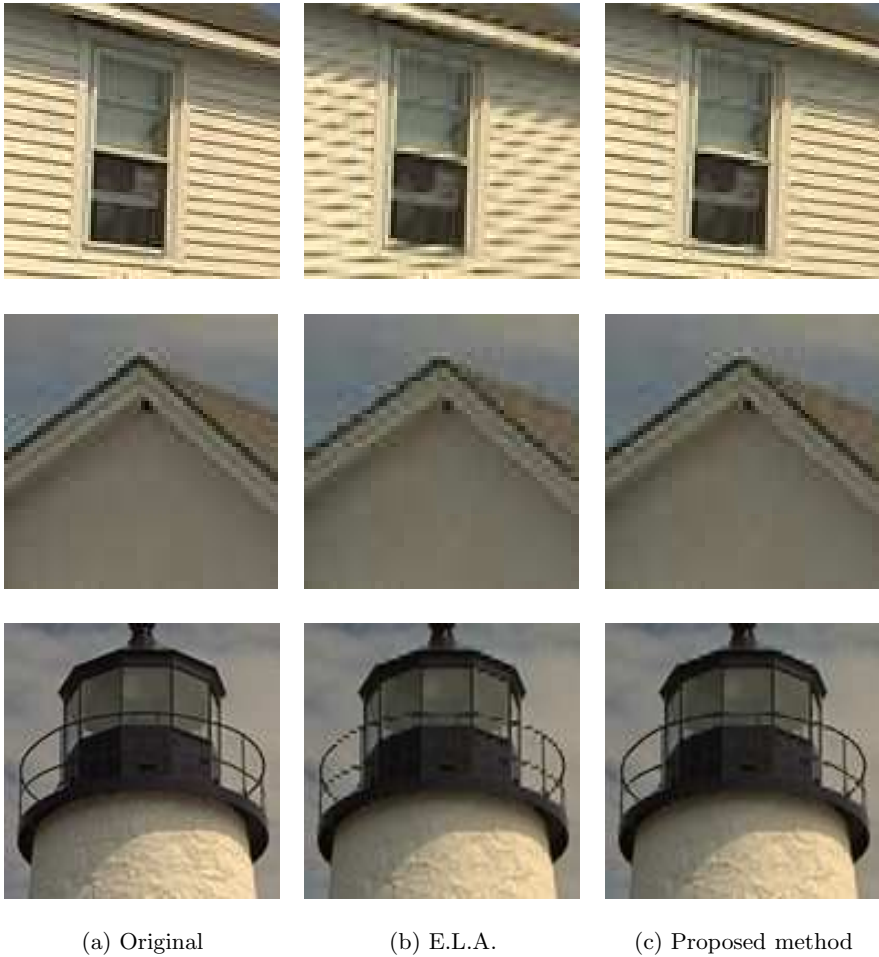


Fig. 11. Comparison of the 2 methods for different parts of the lighthouse image

our method, as described in previous sections, while the E.L.A. method [8] was used to interpolate the other pixels in the image.

If we consider the results obtained on fixed images with many horizontal details, such as the image of the lighthouse (*figure 10*) then we can see that the method reconstructs the continuity of the thin structures in a much better way than the E.L.A. method (*figure 11*). On the moving table tennis sequence, the improvement is also very visible (*figure 12*) already on a still frame. In real time, the benefit of our method is also noticeable since the stability of the moving horizontal lines is assured and the traditional flicker effect of conventional method is removed. The close to horizontal structures which are not properly reconstructed by the traditional methods are almost identical to the structures present in the original progressive material.

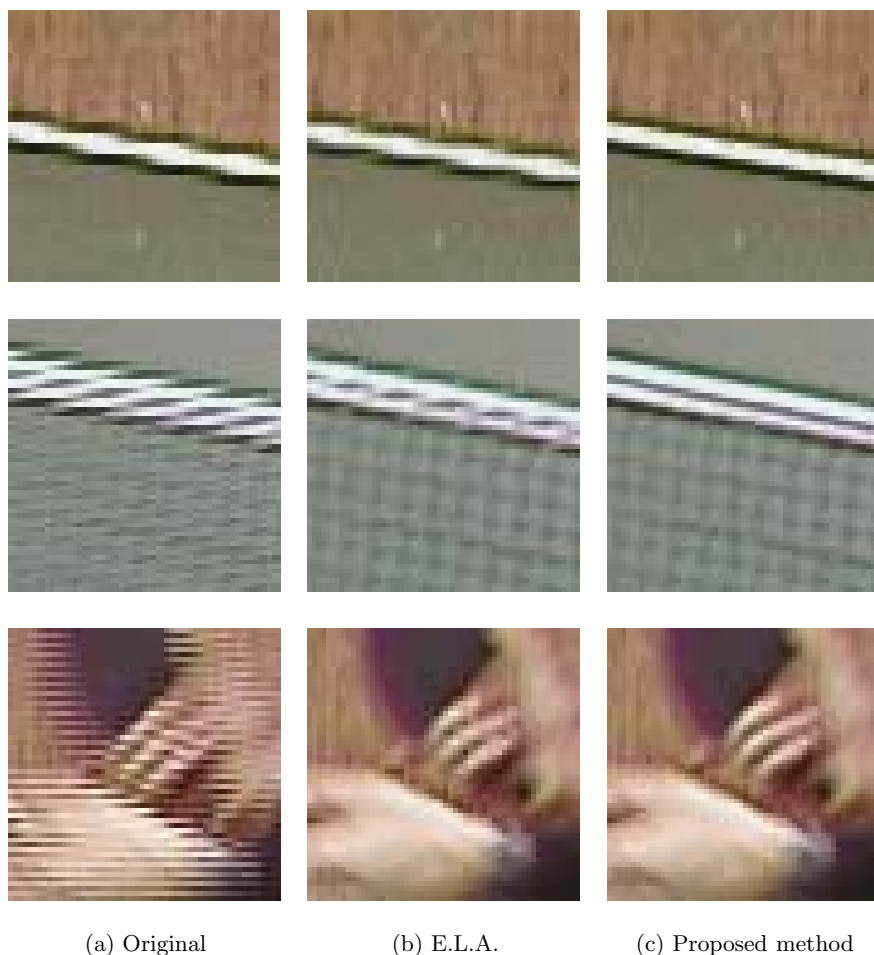


Fig. 12. Comparison of the 2 methods for different parts of the tennis sequence

The processing time hardly increases wrt the processing time of the original E.L.A. method since the method only performs very simple computations and since the amount of chained list data to be analysed is rather small. The added memory required is also very minor wrt to the memory requirements for the rest of a motion adaptive deinterlacing system. We analyzed the number of segments and the number of pixels interpolated by our method for a few sequences (table 1). On average the percentage of pixels interpolated by our method is about 2 % of the whole image. This low percentage is visible in the PSNR comparison (table 2). Indeed, the difference between methods are more significant with the PSNR calculated only on pixels interpolated by our method. Finally, the subjective improvement is noticeable since the eye is very sensitive to the continuity of the linear structures and to their flicker.

Table 1. Number of pixels and segments interpolated by our method for one still image and several video sequences

Sequences	Size	Number of pixels	Extrema pixels		Number of segments	Interpolated pixels	
			Number	% of the image		Number	% of the image
Lighthouse	768×512	393216	15000	3.8	7800	9500	2.4
Car 2	576×720	414720	10000	2.4	3500	5000	1.2
Calendar	576×720	414720	18500	4.5	13000	13000	3.1
BBC	576×720	414720	9500	2.3	3500	7500	1.8
Table Tennis	480×720	345600	9000	2.6	5600	4000	1.2
American banner	480×720	345600	4000	1.2	1900	3600	1.0
Means			11000	2.8	5883	7100	1.8

Table 2. PSNR Comparison (1) PSNR on the whole image (2) PSNR on the pixels interpolated by our method

Sequences	Lighthouse		TableTennis		Starwars		Motor Race	
	(1)	(2)	(1)	(2)	(1)	(2)	(1)	(2)
Line average	30.28	18.17	24.23	17.09	38.04	24.24	30.29	20.31
ELA method	30.23	18.09	24.62	17.96	38.16	24.39	30.54	20.7
Our method	31.48	23.35	25.34	21.16	38.39	26.34	31.14	22.66

9 Conclusion

Our method is not based on the principle of existing methods. It thus does not suffer from the limitations of these methods. The method aims at correcting the most unpleasant artifacts for the human eye by detecting them directly. It is based on the continuity of object limits in order to reconstruct them. So the structures with strong contrast are more stable. Finally, our method can be added to all the traditional methods to improve their weak point without a high added cost. We still have to find an automatic adjustment of the threshold used for the detection of the extrema according to local and global dynamics.

References

1. Sugiyama, K., Nakamura, H.: A method of de-interlacing with motion compensated interpolation. *IEEE Trans. on Consumer Electronics* **45** (1999) 611–616
2. Koivunen, T.: Motion detection of an interlaced video signal. *IEEE Trans. on Consumer Electronics* **40** (1994) 753–760
3. Lin, S.F., Chang, Y.L., Chen, L.G.: Motion adaptive interpolation with horizontal motion detection for deinterlacing. *IEEE Trans. on Consumer Electronics* **49** (2003) 1256–1265
4. de Haan, G., Bellers, E.: Deinterlacing - an overview. *Proceedings of the IEEE* **86** (1998) 1839–1857

5. T.Doyle: Interlaced to sequential conversion for edtv applications. in Proc. 2nd International Workshop Signal Processing of HDTV (1988) 412–430
6. Chen, T., Wu, H.R., Yu, Z.H.: Efficient deinterlacing algorithm using edge-based line average interpolation. *Optical Engineering* **39** (2000) 2101–2105
7. Yoo, H., Jeong, J.: Direction-oriented interpolation and its application to deinterlacing. *IEEE Trans. on Consumer Electronics* **48** (2002) 954–962
8. Park, M.K., Kang, M.G., Nam, K., Oh, S.G.: New edge depend deinterlacing algorithm based on horizontal edge pattern. *IEEE Trans. on Consumer Electronics* **49** (2003) 1508–1512
9. Byun, M., Park, M.K., , Kang, M.G.: Edi-based deinterlacing using edge patterns. *IEEE International Conference on Image Processing* **2** (2005) 1018–1021

Adaptive Macroblock Mode Selection for Reducing the Encoder Complexity in H.264

Donghyung Kim, Jongho Kim, and Jechang Jeong

Department of Electrical and Computer Engineering, Hanyang University,
17 Haengdang, Seongdong, Seoul 133-791, Korea
{kimdh, angel, jjeong}@ece.hanyang.ac.kr

Abstract. The H.264/AVC standard is a video compression standard that was jointly developed by the ITU-T Video Coding Experts Group and the ISO/IEC Motion Picture Experts Group. The H.264 video coding standard uses new coding tools, such as variable block size, quarter-pixel-accuracy motion estimation, intra prediction and a loop filter. Using these coding tools, H.264 achieves significant improvement in coding efficiency compared with existing standards. Encoder complexity, however, also increases tremendously. Among the tools, macroblock mode selection and motion estimation contribute most to total encoder complexity. This paper focuses on complexity reduction in macroblock mode selection. Of the macroblock modes which can be selected, inter8×8 and intra4×4 have the highest complexity. We propose two methods for complexity reduction of inter8×8 and intra4×4 by using the costs of the other macroblock modes. Simulation results show that the proposed methods save about 55% and 74% of total encoding time compared with the H.264 reference implementation when using a full search and a fast motion estimation scheme, respectively, while maintaining comparable PSNR.

1 Introduction

The H.264/AVC standard is a video compression standard that was jointly developed by the ITU-T Video Coding Experts Group and the ISO/IEC Motion Picture Experts Group [1]. To improve coding efficiency, H.264 adopts new coding tools, such as quarter-pixel-accuracy motion estimation (ME), multiple reference frames, a loop filter, variable block size (VBS), etc. [2], [3]. These tools have enabled the standard to achieve higher coding efficiency than prior video coding standards. The encoder complexity, however, increases tremendously.

Several approaches have been proposed to reduce the complexity of the H.264 encoder. Yin et al. proposed a method to alleviate encoder complexity caused by ME and macroblock mode selection [4]. Their low complexity ME algorithm consists of two steps. First, integer-pixel ME is carried out using enhanced prediction zonal search (EPZS). Then, depending on the result of the integer-pixel ME, sub-pixel ME is carried out within some limited areas. To achieve faster macroblock mode selection, their method simply examines limited modes based on the costs of inter16×16, inter8×8, and

inter4×4. Huang et al. proposed an algorithm to reduce the time to search the reference frames for ME complexity reduction [5]. For each macroblock, they analyze the available information after intra prediction and ME from the previous frame to determine whether it is necessary to search more frames. Their method can save about 10-67% of ME computation. Ahmad et al. proposed a fast algorithm for macroblock mode selection based on a 3D recursive search algorithm that takes cost into account as well as the previous frame information [6]. This algorithm leads to a decrease of over 30% in encoding time compared with the H.264 reference implementation. The bitstream length, however, increases by about 15%.

To speed up the H.264 encoding time, we focus on complexity reduction of macroblock mode selection. When an 8×8 DCT is not used, the candidate macroblock modes are SKIP, inter16×16, inter16×8, inter8×16, inter8×8, intra16×16, and intra4×4. An inter8×8 mode can be further partitioned into four sub-macroblock modes: inter8×8, inter8×4, inter4×8, and inter4×4. Among these modes, inter8×8 and intra4×4 modes contribute most to the complexity, especially when rate-distortion optimization (RDO) is used.

In this paper, we propose two algorithms. One is to alleviate inter8×8 complexity. It estimates four sub-macroblock modes within inter8×8 by using the costs of other inter modes with relatively low complexity. The other method reduces intra4×4 complexity, using the similarity between RD costs of two intra modes.

2 Mode Selection Algorithm in the H.264 Reference Software

2.1 Macroblock and Sub-macroblock Modes

The H.264 standard allows the following macroblock modes: SKIP, inter16×16, inter16×8, inter8×16, inter8×8, intra16×16, intra8×8, and intra4×4. Furthermore, each block within inter8×8 can be divided into four sub-macroblock modes. The allowed sub-macroblock modes are inter8×8, inter8×4, inter4×8, and inter4×4. Figures 1 and 2 depict the macroblock partitions of inter and intra macroblock modes, respectively.

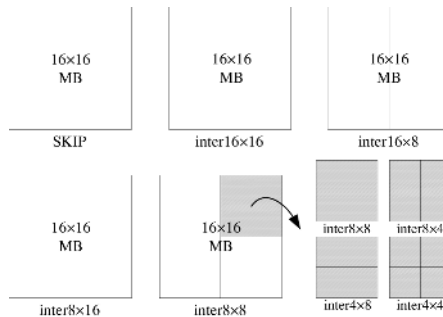


Fig. 1. Macroblock partitions of inter macroblock modes

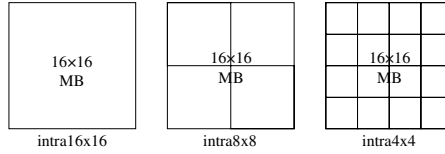


Fig. 2. Macroblock partitions of intra macroblock modes

An inter16×16 mode has only one motion vector, whereas inter16×8 and inter8×16 have two motion vectors. An inter8×8 mode may have 4-16 motion vectors depending on the selected sub-macroblock modes. A SKIP mode refers to the mode where neither motion vector nor residual is encoded. Three intra modes have different prediction modes. Four prediction modes are available in intra 16x16, and nine prediction modes are available in intra8x8 and intra4x4.

2.2 Macroblock Mode Selection in the Reference Software

The reference software, JM9.3 [7], supports three cost calculation criteria: motion vector (MV) cost, reference frame (REF) cost, and rate distortion (RD) cost.

The MV cost is calculated using a lambda factor and is defined as:

$$\begin{aligned}
 \text{MVCost} = & \text{WeightedCost}(f, \text{mvbits}[(cx \ll s) - px] \\
 & + \text{mvbits}[(cy \ll s) - py])
 \end{aligned}$$

where

f : lambda factor
 cx, cy : candidate x and y position for ME
 px, py : predicted x and y position for ME

The REF cost is also calculated using a lambda factor and is defined as:

$$\begin{aligned}
 \text{REFcost} = & \text{WeightedCost}(f, \text{refbits}(\text{ref}))
 \end{aligned}$$

where f : lambda factor

In (1) and (2), *WeightedCost*() returns the cost for the bits of motion vector and reference frame, respectively. Finally, the RD cost is defined as:

$$\begin{aligned}
 \text{RDCost} = & \text{Distortion} + \lambda \cdot \text{Rate}
 \end{aligned}$$

where λ : Lagrange multiplier

In (3), the distortion is computed by calculating the SNR of the block and the rate is calculated by taking into consideration the length of the stream after the last stage of encoding.

When RDO and five reference frames are used, using these cost functions, the process of macroblock mode selection in the reference software is as follows:

- Step 1.** Find reference frames and motion vectors for each block in inter16×16, inter16×8, and inter8×16.

$$[\mathbf{MV}_i, \mathbf{REF}_i] = \arg \min_{MV, REF} (\text{MVcost}(MV) + \text{REFcost}(REF))$$

where $i = \text{inter}16 \times 16, \text{inter}16 \times 8, \text{inter}8 \times 16$.

$MV \in \{\text{Search Range}\}, REF \in \{0, 1, \dots, 4\}$

\mathbf{MV}_i : Motion vector set in i mode

\mathbf{REF}_i : Reference frame set in i mode

(4)

Step 2. Calculate the sums of MV cost and REF cost in $\text{inter}16 \times 16$, $\text{inter}16 \times 8$, and $\text{inter}8 \times 16$.

$$J_i = \text{MVcost}(\mathbf{MV}_i) + \text{REFcost}(\mathbf{REF}_i)$$

where $i = \text{inter}16 \times 16, \text{inter}16 \times 8, \text{inter}8 \times 16$.

J_i : the sum of MV cost and REF cost in i mode.

(5)

Step 3. Find reference frames and motion vectors for the first sub-macroblock in $\text{inter}8 \times 8$.

$$[\mathbf{MV}_i, \mathbf{REF}_i] = \arg \min_{MV, REF} (\text{MVcost}(MV) + \text{REFcost}(REF))$$

where $i = \text{inter}8 \times 8, \text{inter}8 \times 4, \text{inter}4 \times 8, \text{inter}4 \times 4$.

(6)

Step 4. Calculate the sums of MV cost and REF cost for the first sub-macroblock in $\text{inter}8 \times 8$.

$$J_i = \text{MVcost}(\mathbf{MV}_i) + \text{REFcost}(\mathbf{REF}_i)$$

where $i = \text{inter}8 \times 8, \text{inter}8 \times 4, \text{inter}4 \times 8, \text{inter}4 \times 4$.

(7)

Step 5. Select the mode for the first sub-macroblock in $\text{inter}8 \times 8$.

Sub-macroblock mode =

$$\arg \min_{mode} (\text{RDcost}(\text{inter}8 \times 8), \text{RDcost}(\text{inter}8 \times 4)$$

$$, \text{RDcost}(\text{inter}4 \times 8), \text{RDcost}(\text{inter}4 \times 4))$$

(8)

Step 6. Repeat steps 3 to 5 for the other sub-macroblocks in $\text{inter}8 \times 8$.

Step 7. Select the macroblock mode

Macroblock mode =

$$\arg \min_{mode} (\text{RDcost}(SKIP), \text{RDcost}(\text{inter}16 \times 16)$$

$$, \text{RDcost}(\text{inter}16 \times 8), \text{RDcost}(\text{inter}8 \times 16), \text{RDcost}(\text{inter}8 \times 8)$$

$$, \text{RDcost}(\text{intra}16 \times 16), \text{RDcost}(\text{intra}4 \times 4))$$

(9)

In steps 1 and 2, the reference software finds reference frames and motion vectors which minimize the sum of MV cost and REF cost in $\text{inter}16 \times 16$, $\text{inter}16 \times 8$, and $\text{inter}8 \times 16$. Steps 3 to 6 are the process of selecting sub-macroblock modes in $\text{inter}8 \times 8$. The final step decides the macroblock mode by comparing RD costs of all macroblock modes.

3 Proposed Algorithm

3.1 Complexity Reduction of Inter8x8

Since each sub-macroblock within inter8x8 needs additional RD cost computations for the selection of sub-macroblock modes, inter8x8 has the highest complexity among all of the inter macroblock modes. For complexity reduction of inter8x8, we assume that the costs of inter macroblock modes monotonically increase or decrease according to their partitioned direction. Under this assumption, we restrict selectable sub-macroblock modes by using the MV costs and REF costs of inter16x16, inter16x8, and inter8x16. For example, if the sum of MV and REF costs of inter16x16 is larger than that of inter16x8 and is smaller than that of inter8x16, we consider only inter8x8 and inter8x4 as sub-macroblock modes. Figure 3 depicts the proposed method for the complexity reduction of inter8x8.

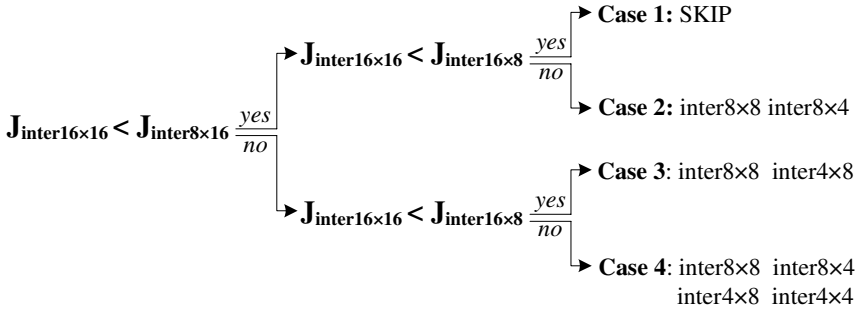


Fig. 3. Block diagram for the restriction of selectable sub-macroblock modes

In case 1, since $J_{inter16x16}$ is smaller than both $J_{inter16x8}$ and $J_{inter8x16}$, neither additional block partition in the horizontal direction nor in the vertical direction is needed. In this case we do not consider any sub-macroblock mode, and step 3 to step 6 in the reference software are skipped. In case 2, since $J_{inter16x16}$ is smaller than $J_{inter8x16}$ and is larger than $J_{inter16x8}$, additional block partitioning is only considered in the vertical direction. In this case, either inter8x8 and inter8x4 is selected as a sub-macroblock mode, and the formulae of steps 3 to 6 in the reference software are modified as follows:

$$[\mathbf{MV}_i, \mathbf{REF}_i] = \arg \min_{\mathbf{MV}, \mathbf{REF}} (\mathbf{MVcost}(\mathbf{MV}) + \mathbf{REFcost}(\mathbf{REF})) \tag{10}$$

where $i = inter8x8, inter4x8$.

$$J_i = \mathbf{MVcost}(\mathbf{MV}_i) + \mathbf{REFcost}(\mathbf{REF}_i) \tag{11}$$

where $i = inter8x8, inter8x4$.

Sub - macroblock mode =

$$\arg \min_{mode} (\mathbf{RDcost}(inter8x8), \mathbf{RDcost}(inter8x4)) \tag{12}$$

In case 3, since $J_{\text{inter}16 \times 16}$ is smaller than $J_{\text{inter}16 \times 8}$ and is larger than $J_{\text{inter}8 \times 16}$, only additional block partition is considered, and only in the horizontal direction. In this case, either $\text{inter}8 \times 8$ and $\text{inter}4 \times 8$ is selected as a sub-macroblock mode, and the formulae of steps 3 to 6 in the reference software are modified as follows:

$$[\mathbf{MV}_i, \mathbf{REF}_i] = \underset{MV, REF}{\text{arg min}} (\text{MVCost}(MV) + \text{REFCost}(REF)) \quad (13)$$

where $i = \text{inter}8 \times 8, \text{inter}4 \times 8$.

$$J_i = \text{MVCost}(\mathbf{MV}_i) + \text{REFCost}(\mathbf{REF}_i) \quad (14)$$

where $i = \text{inter}8 \times 8, \text{inter}4 \times 8$.

$$\begin{aligned} \text{Sub-macroblock mode} = \\ \underset{\text{mode}}{\text{argmin}} (\text{RDcost}(\text{inter}8 \times 8), \text{RDcost}(\text{inter}4 \times 8)) \end{aligned} \quad (15)$$

In case 4, since $J_{\text{inter}16 \times 16}$ is larger than both $J_{\text{inter}16 \times 8}$ and $J_{\text{inter}8 \times 16}$, we consider all sub-macroblock modes, as in the reference software.

3.2 Complexity Reduction of Intra4×4

When 8×8 DCT is not used, the allowed intra modes are $\text{intra}4 \times 4$ and $\text{intra}16 \times 16$. Of the two intra modes, $\text{intra}4 \times 4$ has the higher complexity because it has more prediction modes. Since $\text{intra}16 \times 16$, as described in Section 2, has only four prediction modes and $\text{intra}4 \times 4$ has nine prediction modes for finer granularity, $\text{intra}4 \times 4$ generally yields a smaller prediction error than $\text{intra}16 \times 16$. However, most of the macroblocks have only a small difference between the RD costs of $\text{intra}16 \times 16$ and $\text{intra}4 \times 4$. This is because edges directed in vertical or horizontal directions are dominant in natural images, which are considered in $\text{intra}16 \times 16$.

Using this characteristic, we first find the inter mode with a minimum RD cost. Then we compare the RD cost of the selected inter mode with that of $\text{intra}16 \times 16$. If the RD cost of $\text{intra}16 \times 16$ is much larger, that is, if Eq. (16) is true, then the RD cost computation of $\text{intra}4 \times 4$ is skipped:

$$\text{Min}[\text{RDcost}(\text{inter modes})] \cdot K < \text{RDcost}(\text{intra } 16 \times 16) \quad (16)$$

In (16), K is a proportional constant. Table 1 describes the missing rate of $\text{intra}4 \times 4$. The missing rate indicates the probability that the skipped $\text{intra}4 \times 4$ has the smallest RD cost. As shown in Table 1, the average missing rate is only about 0.7% for $K = 1.5$. This means that the RD cost difference factor between $\text{intra}4 \times 4$ and $\text{intra}16 \times 16$ is less than 1.5 for 99.3% of the macroblocks.

4 Simulation Results

Since the proposed methods for complexity reduction of $\text{inter}8 \times 8$ and $\text{intra}4 \times 4$ are uncorrelated, the two methods can be applied independently or simultaneously. We applied the two proposed algorithms simultaneously to encode test sequences. For the purpose of evaluation, the public reference encoder JVT Model (JM) v.9.3 was used. The software was tested on an Intel Pentium-IV based computer with 1024 MB RAM under the Windows XP Professional operating system.

Table 1. Missing rates of intra4x4 according to the proportional constant K

Sequences	Missing Rate (%)		
	$K=1.3$	$K=1.5$	$K=1.7$
Coastguard	2.0	0.4	0.2
Container	2.7	1.0	0.9
Mobile	0.4	0.0	0.0
News	5.8	0.6	0.5
Salesman	6.6	0.4	0.0
Silent	4.7	1.7	0.4
Stefan	2.8	0.2	0.0
Trevor	9.0	0.9	0.1

We adopted two different schemes for ME, used RDO, and set Quantization Parameter (QP) and K in (16) to 28 and 1.5, respectively. The simulation was performed on eight standard video sequences in QCIF (176x144) format. These included Coastguard, Container, Mobile, News, Salesman, Silent, Stefan, and Trevor. These sequences were selected on the basis of length of encoded streams and degree of motion. The first 100 frames of each of these sequences were used.

For 99 P-frames, Tables 2 and 3 describe the reduction ratios of the number of RD cost computations in inter8x8 and intra4x4 in case when using a full search and a fast motion estimation scheme. As shown in these results, we can save about 72% and 89% of the RD cost computations, respectively.

Tables 4 and 5 compare the bitrates and PSNRs for each test sequence. Since the reference implementation is an exhaustive search for selecting the macroblock mode, the number of encoded bits is the least for each sequence. Tables 4 and 5 show the average increase of the total bitrates is only about 0.9%, and the average PSNR drop is only about 0.044 dB when using the proposed method.

Finally, Table 6 compares total encoding time from the proposed method with that from the reference software. This result shows a substantial decrease of about 55% and 74% in total encoding time compared with the reference implementation when adopting a full search and a fast motion estimation algorithm, respectively.

Table 2. The number of RD cost computation in inter8x8 when using (a) a full search (FS) motion estimation scheme (b) a fast motion estimation (FME) scheme for luminance blocks

Sequences	Reference Software	Proposed Method (a)	Proposed Method (b)	Average Reduction (%)
Coastguard	156,816	59,760	59,768	61.9
Container	156,816	17,896	18,072	88.5
Mobile	156,816	65,360	65,280	58.3
News	156,816	30,256	28,752	81.2
Salesman	156,816	28,224	28,208	82.0
Silent	156,816	40,464	40,120	74.3
Stefan	156,816	56,472	56,952	63.8
Trevor	156,816	54,800	54,904	65.0
Average	156,816	44,154	44,007	71.9

Table 3. The number of RD cost computation in intra4x4 when using (a) a full search (FS) motion estimation scheme (b) a fast motion estimation (FME) scheme for luminance blocks

Sequences	Reference Software	Proposed Method (a)	Proposed Method (b)	Average Reduction (%)
Coastguard	9,801	2,907	2,864	70.6
Container	9,801	1,993	2,086	79.2
Mobile	9,801	122	130	98.7
News	9,801	655	669	93.2
Salesman	9,801	145	159	98.4
Silent	9,801	637	633	93.5
Stefan	9,801	867	953	90.7
Trevor	9,801	995	1,040	89.6
Average	9,801	1,040	1,067	89.3

Table 4. Bitrates (Kbits/sec) of test sequences when using (a) a full search (FS) motion estimation scheme (b) a fast motion estimation (FME) scheme for luminance blocks

Sequences	Reference Software	Proposed Method (a)	Proposed Method (b)	Average Increase (%)
Coastguard	249.00	251.28	249.96	0.7
Container	40.16	40.74	40.57	1.2
Mobile	496.49	497.24	497.65	0.2
News	75.84	76.75	76.27	0.9
Salesman	56.89	57.61	57.74	1.4
Silent	82.69	83.71	84.33	1.6
Stefan	379.26	380.86	380.62	0.4
Trevor	132.49	133.37	133.91	0.9
Average	189.10	190.19	190.13	0.9

Table 5. PSNRs (dB) of test sequences when using (a) a full search (FS) motion estimation scheme (b) a fast motion estimation (FME) scheme for luminance blocks

Sequences	Reference Software	Proposed Method (a)	Proposed Method (b)	Average PSNR Decrease
Coastguard	33.93	33.89	33.88	0.045
Container	36.07	36.06	36.05	0.015
Mobile	33.14	33.06	33.05	0.085
News	36.65	36.64	36.66	0.000
Salesman	35.57	35.54	35.54	0.030
Silent	35.84	35.81	35.81	0.030
Stefan	34.22	34.15	34.12	0.085
Trevor	36.40	36.34	36.34	0.060
Average	35.23	35.19	35.18	0.044

Table 6. Encoding time (sec) of test sequences when using (a) a full search (FS) motion estimation scheme (b) a fast motion estimation (FME) scheme for luminance blocks

Sequences	Reference Software	Proposed Method (a)	Reduction Ratio (%)	Proposed Method (b)	Reduction Ratio (%)
Coastguard	127.64	70.08	45.10	45.08	64.68
Container	110.77	51.19	53.79	27.62	75.06
Mobile	142.67	57.89	59.42	35.77	74.93
News	112.59	49.06	56.43	26.46	76.50
Salesman	115.64	47.06	59.31	24.19	79.09
Silent	113.77	51.52	54.72	28.59	74.87
Stefan	130.56	58.30	55.35	36.36	72.15
Trevor	110.35	54.15	50.93	31.27	71.67
Average	120.50	54.91	54.38	31.92	73.62

5 Conclusions

The H.264 video coding standard uses new coding tools. Among the tools, the macroblock mode selection process tremendously increases the encoder complexity. For reducing the encoder complexity, we proposed two simple and effective schemes for the quick selection of macroblock modes in H.264 video coding.

Using our methods, the RD cost computations of inter 8×8 and intra 4×4 were reduced by about 72% and 89%, respectively. Both schemes can be applied independently. When both methods are used simultaneously and two different motion estimation methods are applied, simulation results show that our methods can save about 55% and 74% of total encoding time regardless of input sequences, respectively, yet the average increased rate of the total bits and average PSNR drop are only about 0.9% and 0.044 dB, respectively. This huge reduction of encoder complexity may be useful in real-time implementation of the H.264/AVC standard.

Acknowledgement

This work was supported by the Ministry of Commerce, Industry and Energy with the project of Development of Personal Next Generation TV terminal.

References

1. Wiegand, T.: Version 3 of H.264/AVC. Doc. JVT-K051 (2004)
2. Wiegand, T., Sullivan, G. J.: Overview of the H.264/AVC Video Coding Standard. IEEE Trans. Circuits Syst. for Video Technol., Vol. 13. (2003) 560-576
3. Wiegand, T., Schwarz, H., Joch, A., Kossentini, F.: Rate-Constrained Coder Control and Comparison of Video Coding Standard. IEEE Trans. Circuits Syst. for Video Technol., Vol. 13. (2003) 688-703
4. Yin, P., Tourapis, H. C., Tourapis, A. M., Boyce, J.: Fast Mode Decision and Motion Estimation for JVT/H.264. ICIP'03, Vol. 3. (2003) 853-856

5. Huang, Y. W., Hsieh, B. Y., Whang, T. C., Chien, S. Y., Ma, S. Y., Shen, C. F., Chen, L. G.: Analysis and Reduction of Reference Frames for Motion Estimation in MPEG-4 AVC/JVT/H.264. ICASSP'03, Vol. 3 (2003) 145-148
6. Ahmad, A., Khan, N., Masud, S., Maud, M.A.: Efficient Block Size Selection in H.264 Video Coding Standard. Electronics Letters, Vol. 40. (2004) 19-21
7. JM9.3: <http://bs.hhi.de/~suehring/tml/download/jm93.zip>.

Dynamic Light Field Compression Using Shared Fields and Region Blocks for Streaming Service

Yebin Liu, Qionghai Dai, Wenli Xu, and Zhihong Liao

Department of Automation, Tsinghua University, Beijing 100084, P.R. China
liuyb02@mails.tsinghua.edu.cn

Abstract. The multi-view simul-switching is one of the most important features of dynamic light field (DLF) streaming. In this paper, we jointly consider light field rendering and compression and propose a novel DLF compression scheme based on the requirements of DLF streaming. In this scheme, successive temporal prediction chains are broken and a shared field is used as a reference for all the later P frames in a group of field. Meanwhile, considering the region of interest for DLF rendering, we partition all the P frames into regional blocks and code them in a manner that any region block can be independently transmitted for bandwidth economized streaming. With this coding scheme, a multi-camera DLF system is developed to verify the streaming performance of the proposed scheme. Experimental results show that our scheme saves the per-user transmission bandwidth a lot compared with other DLF compression schemes.

1 Introduction

The ability to interactively and seamlessly roam in a scenario while watching a streaming video through IP network is an exciting visual experience. Thanks to Levoy[1] and Gortler's[2] publication of light field rendering technique, this experience is no longer a dream. A light field is a collection of light rays following through space in all directions captured by a multi-camera array and recorded as multi-view images which allow seamless view generation with only a little geometry information or even none geometry information involved. The introduction of temporal dimension to light field produce a new media called dynamic light field(DLF) [3] or plenoptic video[4]. With the advent of this new media and its corresponding multi-camera array technique, just-in-time capturing and rendering of dynamic scene is guaranteed. To verify such real-time interactive rendering ability for real dynamic scenes and to improve the rendering performance, lots of DLF environments have been developed in recent years. B.Wilburn [5] has implemented an MPEG2 light field camera array to capture and store DLF. J.C.Yang [6] has developed a light field rendering system that can interactively render 3D scene. T.Naemura [7] constructed a camera array system consisting of 16 cameras for real-time rendering using multiply focal planes.

Despite the above works on DLF system and technologies, the fundamental issues for DLF to become a popular media have not been investigated enough,

especially in an IP network streaming scenario. It is well known that large data amount is one of the challenges to DLF transmission and storage. In the past decade, there is still little work on DLF compression while a number of multi-view video coding (MVC) techniques have been proposed [8,9,10]. These schemes usually employ a temporal-spatial prediction manner to deal with the overall compression efficiency. However, the key characteristic of DLF streaming and multi-view video streaming lies in the interactivity, which can give users the opportunity to choose their favorite camera-streams freely. The successive temporal predictions used in these schemes block the random switching between views, thus to guarantee just-in-time interactivity, all the camera-streams must be transmitted which is not practical in real streaming systems. Recently, two multi-view video compression schemes offer the feature of free view-point switching are reported in [11] and [12]. Both of these schemes employ a similar SP frame [13] mechanism by using global motion estimation (GME) or wyner-ziv frames and they may still introduce reconstruction error propagation when switching happens. Actually, since the purpose of DLF steaming and multi-view video streaming are different (for DLF streaming, user may choose multiple views but not necessary the integral region of the views to render the new image, while for conventional multi-view streaming, user may choose only one integral view), all the above coding schemes are not well fit for DLF streaming.

Toward the goal of DLF compression for network streaming, in this paper, we present a DLF compression scheme based on the nature of DLF rendering and the requirements of DLF streaming. First, successive temporal prediction chains are broken and a shared field is used as a reference for all the later P frames in a group of field. Therefore, once the shared field is correctly received, users can switch freely between camera-streams. Second, considering the region of interest for rendering, we segment all the P frames into several regional blocks(RBs) and code them in a manner that any region blocks can be independently decoded without any other region blocks. Thus the server can selectively transmit the contents within each camera-stream according to the user requirements and network bandwidth saved.

The rest of the paper is organized as follows. Section 2 presents the general DLF streaming system and explains the major challenges in such application system. Section 3 describes the details of our DLF compression scheme. In section 4, we report and analyze our experiment results. Finally, conclusions are drawn in section 5.

2 System Architecture and Problem Statement

Figure 1 illustrates the flow of a typical DLF streaming system. The multiple camera-views are encoded using some DLF compression scheme. For the forward channel, the bitstreams necessary for new image rendering are transmitted from the server, and the client receives these bits and decode them for rendering operation. As for the backward channel, the user at the client can freely select the preferred streams and send the selections to the server. Then the server



Fig. 1. Typical DLF streaming system

must send the new streams corresponding to the selection of the client. Error free decoding of the new streams must be guaranteed throughout the streaming process.

Since the server bears only data streaming, the above server system serves for multiple users can be practically realized. Users at the clients can enjoy seamlessly roaming in a scenario through the following three kinds of view trajectory:

- 1) Time frozen movement: Users can choose to have a pause and roam in the scene for the interested people or object smoothly or abruptly.
- 2) Time continual movement: Users are able to change the viewing position and viewing direction as the video continues along time.
- 3) View zooming: DLF rendering can provide zooming capability for users. If the user trajectories are near the camera plane, a relatively small number of frames are required to generate new views. Otherwise, as for the situation of zooming in and out from the camera plane, more frames are needed.

As stated above, one of the main challenges of such DLF streaming service lies in the just-in-time interactivity. Traditional successive temporal prediction may be limited to key frame switching as illustrated in figure 2. Before instant t_3 , the views user demanded and sever sent are V_1 and V_2 . But when t_3 comes, the server receives the requirement of view V_3, V_4 and V_5 . Usually, it happens to be a P frames at t_3 , and thus the server must wait for the next key frames and timely response is destroyed. Because of the simultaneously switching of multiple views, we called such situation as simul-switching. Although some SP-frame-like information can be added to force timely switching [11,12], they will introduce additional storage and error drifting problem.

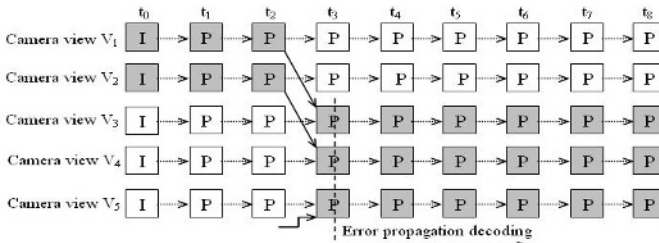


Fig. 2. Switching problem in DLF streaming system

3 Proposed DLF Compression Scheme

In this section, our implementation of DLF compression based on hybrid video codec for streaming service will be described. A shared I field coding scheme is presented in 3.1. In 3.2, we analysis the region of interest for light field rendering and then in 3.3, based on the freely switching coding scheme in 3.1, the idea of independent region blocks compression is incorporated to minimize per-user streaming bit rate. In 3.4, we will give the corresponding streaming policy for our coding scheme.

3.1 Shared I Field Coding Scheme

Figure 3 represents the shared I field coding scheme. In the DLF, the images captured at the same instant constitute a static light field which is similar to a "frame" in video coding. We define 2 types of light fields, and they are I field and P field. A GOF (group of fields) is composed of an I field and all its following P fields. The traditional temporal prediction chains can be break and successive prediction correlation will be eliminated using the corresponding image in the former I field as the reference image for each image in P fields and the later I fields. The compression efficiency of such prediction is still high since the camera array and the background of the scene are both static. As for data streaming, I field is imperative for every clients while images in P field can be selectively transmitted. Therefore, once I field is successfully received, images in P field can be freely transmitted without the consideration of switching. Meanwhile, correctly and just-in-time decoding is achieved.

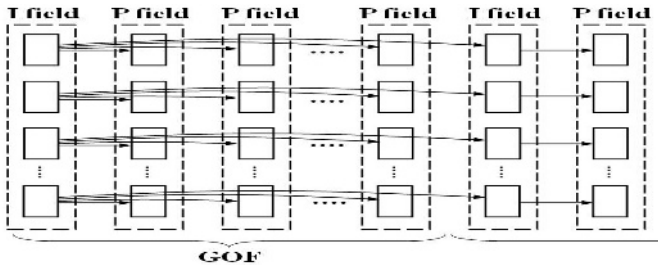


Fig. 3. Prediction structure for the DLF compression

Certainly, we can improve the compression efficiency through the using of several or even all of the images in I fields for multi-hypothesis prediction coding of each image in P fields. In addition, spatial layered prediction [14] or disparity-compensated prediction [15] can be employed to improve the compression efficiency of the I fields. However, these two operations may introduce data exchange between camera-views which complicate the coding operations.

3.2 Region of Interest Analysis in Light Field Rendering

There are different rendering schemes [1,2,16,17] (with or without geometry involved, regular cameras or unstructured cameras) in the literature, and all of these schemes have the region selection (from the available camera views) and the region blending process. Here, we use dynamically reparameterized light fields rendering [16] to illustrate the region selection procedure. First, the data camera's aperture filter is projected on to the virtual camera's(or desired camera's) image plane producing the region which uses samples from the data camera. Then, the data camera's aperture filter is projected on to the focal plane generating the viewing content from this data camera. Such viewing content on the focal plane is then re-projected on the data camera plane from the data camera's point of view. This projection produces the interested region on the data camera's image. At last, the interested region is texture mapped on to the desired image plane's region which has been computed in the first step. Multiply texture mappings from all the single data cameras may output the final desired image.

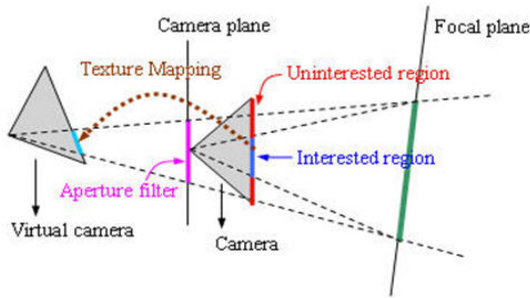


Fig. 4. Illustration of general light field rendering

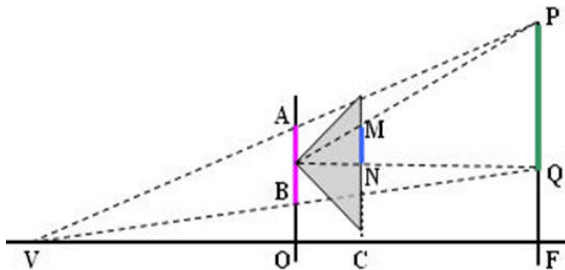


Fig. 5. Geometry in light field rendering when the focal plane is parallel to the camera plane

A more straightforward illumination of the region-of-interest for light field rendering is shown in figure 5. Under the presuppositions that the capture range of each data camera is broad enough and the focal plane is parallel to the camera plane, the following equations can be established from this figure:

$$|MN|/|PQ| = |OC|/|OF| \quad (1)$$

and

$$|PQ|/|AB| = |VF|/|VO|. \quad (2)$$

Therefore, the region of interest $|MN|$ can be derived as:

$$|MN| = |AB||OC||VF|/(|OF||VO|). \quad (3)$$

Based on the figure and equations above, several conclusions concerned with the region of interest for rendering can be made as follows:

Conclusion 1: If the desired view range is unlimited, the region of interest for the data camera is irrelevant to the desired view direction but relevant to the view position. The closer the view position to the camera plane, the broader the interested region will be.

Conclusion 2: The longer the distance is between the camera plane to the focal plane, the more narrow the interested region for a particular camera will be, but the larger the number of data cameras that must contribute interested regions to the rendering process.

3.3 Independent Region Block Coding for P Pictures

In our scheme, an image can be completely and regularly partitioned by several region blocks which is composed of some macroblocks. As for a P-field image at 320×240 resolution, there are 24 partition modes and the corresponding region blocks can be of size $16m \times 16n$ ($m = 1, 2, 4, 5, 10, 20, n = 1, 3, 5, 15$). Figure 6 illustrates the partition of region block at size 80×80 . The required area for streaming under a particular mode is the minimal set of region blocks that can cover the region of interest.

Each region block is independently coded similar to the slice partition mechanism in H.264/AVC. However, unlike slice partition, our block partition achieves random access operation to any region block when it is combined with the shared I field coding mechanism. Through the decorrelation of motion vectors on the edge of region blocks and the insertion of synchronization bits at the start of each region block's code bits, all data in P fields are coded as independent region block streams for bandwidth economized streaming application. Once the partition mode and the camera parameters have been determined, the required region block streams for the particular virtual view rendering can be computed and recorded as look-up table beforehand using the model in 3.2. It must be noted that finely partitioned may results in economized area for transmission but worse compression ratio. Hence, the choice of region block size must be based on rendering algorithm and coding characteristic.

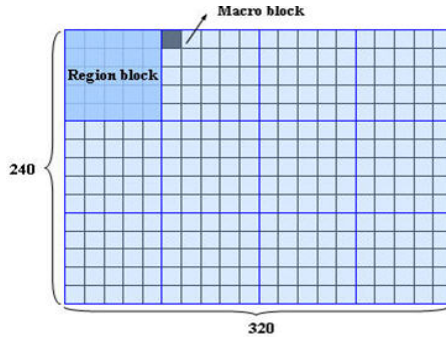


Fig. 6. 80×80 region block partition for SIF format image

3.4 DLF Streaming

For compressed DLF streaming, the compressed DLF are stored in the storage device in a manner that all the region block streams are independent between each other. Once a client requests for streaming service, the streaming server reads all the region block streams into the memory and buffers them for several seconds (usually within 2 seconds). As an user changes his viewpoint or view direction, the client checks in the look-up table and obtains the region blocks needed for rendering and sends back the request through the feedback channel to the streaming server. Once the streaming server receives the message, it switches the region blocks streams and sent them through the data channel. As for I fields, each user receives a full copy of the stream and decodes all the I images that are captured at the same instant for region blocks decoding. If the speed of data decoding and image rendering is fast enough, the controlling delay user experienced may approximates to only the time of request message feedback.

4 Experiment Results

We use 64 BOSER BS-103F color cameras 8 bits per pixel CCD sensors to set up an 8×8 light field camera array (see figure 7) and capture a 10 second 4×8 dynamic light field sequences with image size of 320×240 and frame rate 30fps. The optical axis of each camera is roughly perpendicular to a common camera plane. The horizontal spacing between cameras is about 8cm, and the vertical spacing is about 14cm. The cameras are connected by IEEE-1394 High Performance Serial Bus to the producer PCs. Every 4 cameras (the one camera with its right, bottom and diagonal neighbored cameras) are connected to one of the 8 producer PCs. Figure 8 shows the first light field of this "Room2" sequences. We implement our shared I fields and region blocks based coding scheme through the modification of the Mpeg4 XVID codec.

First, we examine the shared I field prediction efficiency. Figure 9 illustrates the intra compression efficiency, temporal prediction efficiency, spatial prediction



Fig. 7. A photo of our 64-camera light field camera array. The cameras are arranged in rows of eight.



Fig. 8. 4×8 views in the first field of DLF sequence taken with our light field camera

efficiency and I field prediction efficiency under the same quantization configuration. Here, all prediction codings are implemented using only one reference frame. Although the color calibration is satisfactory, the spatial prediction efficiency is still much lower than temporal prediction. For our proposed coding scheme, with the increasing interval between I field and the coding frame, the prediction efficiency becomes lower but still shows better performance than spatial prediction efficiency and approximate to conventional video coding. Therefore, the abandonment of spatial prediction in our coding scheme for convenient purpose is reasonable.

Second, we examine the coding performance of the shared I field compression when region blocks coding mechanism is introduced. Figure 10 depicts the rate distortion characteristics of one of the views in our Room2 DLF sequence under conventional XVID video coding scheme and our region block coding scheme.

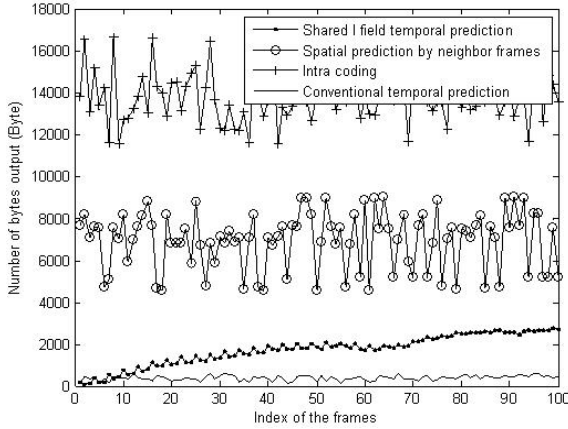


Fig. 9. Comparison of prediction efficiency

According to the figure, we see that when the shared length (length of GOF or I field interval) is 30 without region partition (RB size = 320×240), our coding performance is comparable to tradition video coding with key frame interval set to be 15. As for the region partition, when the image is partitioned into 6 region blocks with size 80×80 , the increased bit-rate is only about 3kbps compared with the case of no partition under the approximate reconstruction image quality. However, further partition with region block size 32×48 , the increased bit rate will be about 25kbps and when finely partition is achieved, the coding performance will drop rapidly as the situation of RB size 16×16 . Actually, the additional bits introduced by independent motion vector coding is trivial compared with the increasing of synchronization bits which grows up linearly with the number of region blocks.

Third, we test the average saved area under various region partition modes. Figure 11 depicts the interested region for our 4×8 camera array light field rendering. Each block in this figure represents a view (at size 320×240) while the black regions in the blocks represent the regions of interest. We can actually feel that our region blocks coding have the potentiality to save the transmission bandwidth. Also, an interactive rendering viewer is developed and a 10 minutes long view roaming is simulated using this viewer. The view trajectories are recorded and then the average percentages of image area saved (the sum of unnecessary region areas in the required images divided by the sum of required image areas) for streaming are computed under various partition modes. Table 1 lists the results for the 24 kinds of partition modes. The elements in the first column of the table are the 4 heights of the region block while the elements in the first line are the 6 widths. Note that the saved percentage increases with the decreasing of size of the region block sizes, but such uptrend tends to be gentle when the region block size arrives 80×80 . Also, it is shown that quadrature region block is better because of the isotropic region requirement at both horizontal direction

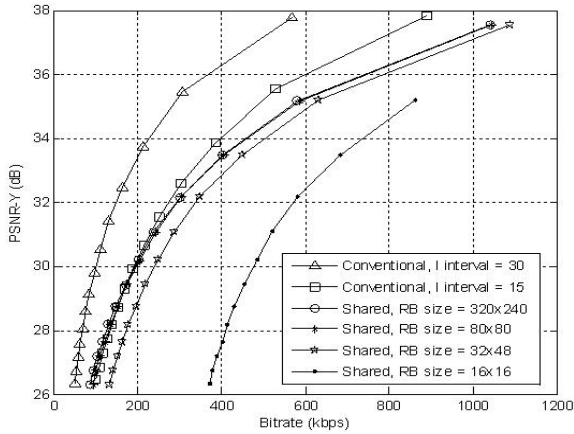


Fig. 10. Coding performance under different coding schemes. For the four shared I field coding scheme, the GOF length is 30.

and vertical direction for planer light field rendering. Furthermore, since the coding performance impairment of partition mode 80×80 is imperceptible as has been verified in figure 10, we choose this mode for our DLF compression.

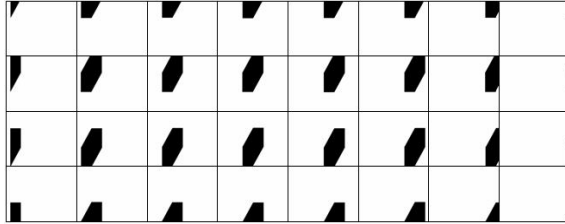


Fig. 11. An example of region of interest in the 4×8 light field

At last, we investigate the streaming performance of our coding scheme using our 4×8 DLF sequence. Figure 12 compares the average streaming bit rates under various coding schemes. We fix the average image coding quality within the range from 35.8dB to 36.1dB and restrict the distance between the user view position and the camera plane to several constant values. At each value, the user can still rooming in the scene by changing his view point, view direction and focal plane. For the traditional video coding scheme, all the 32 views must all be streamed for timely switching. From the figure we can see that the streaming bit rate of our proposed coding scheme maintains constant with the changing of distance and obtains the minimum transmission bit rate. Also, when the view point is near the camera plane, the required bandwidth is smaller than 2Mbps.

Table 1. Average percentages of saved area under various region block sizes

RB size	320	160	80	64	32	16
240	0%	25%	36.7%	38.8%	46.6%	49.1%
80	48.9%	62.0%	68.2%	69.6%	73.2%	74.9%
48	62.5%	71.8%	76.4%	77.3%	79.9%	80.8%
16	67.7%	76.4%	80.7%	81.6%	83.7%	84.7%

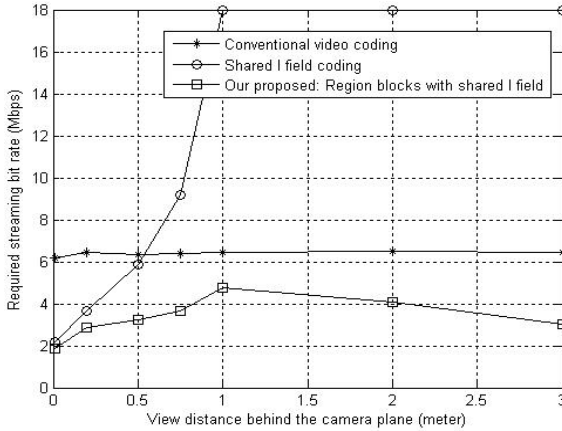


Fig. 12. Average required transmission bandwidth changing with the distance between the desired view point and the camera plane. The vision quality is set to be about 36dB.

5 Conclusions

In this paper, we have advanced the DLF streaming service and described the main technical challenges for this wonderful application, namely the tremendous data amount and the error free multiple view simul-switching problem. Unlike previous algorithms for multi-view video (MVC) or DLF coding, we jointly consider the coding and the rendering procedure and design a region block based compression scheme for bandwidth economized streaming. Another specialty of our work is that this region block based mechanism can be perfectly combined with our shared I field coding scheme to generate the switching permit DLF coding scheme. From the experimental results, we have observed that our compression framework realizes DLF streaming under 2Mbps bit rate when the user’s view point is near the camera plane. Such bandwidth requirement is suitable for broadband IP network.

Acknowledgment

This work is supported by the Distinguished Young Scholars of NSFC(No. 60525111) and the key project of NSFC(No.60432030)

References

1. M. Levoy and P. Hanrahan.: Light field rendering. In: Computer Graphics (Proceedings. SIGGRAPH96), pages 31-42, August 1996
2. S. J. Gortler, R. Grzeszczuk, R. Szeliski, and M. F. Cohen.: The lumigraph. In: Computer Graphics (Proceedings SIGGRAPH-96), pages 31-42, Aug. 1996
3. B. Goldlucke, M. Magnor, B. Wilburn.: Hardware-accelerated Dynamic Light Field Rendering. In: Proceedings of Vision, Modeling, and Visualization 2002, VMV2002, Germany, November 20-22
4. S.C.Chan, K.T.Ng, Z.F.Gan, K.L.Chan, and H.Y.Shum.: The plenoptic video. In: IEEE Trans on Circuits and Systems for Video Technology, vol. 15, no. 12, pp. 1650-1659, Dec 2005
5. B. Wilburn, M. Smulski, K. Lee, and M. Horowitz.: The light field video camera. In: Proc. SPIE Electronic Imaging: Media Processors, vol. 4674, Jan. 2002, pp. 29-36
6. Jason C. Yang, Matthew Everett, Chris Buehler, Leonard McMillan.: A Real-Time Distributed Light Field Camera. In: Thirteenth Eurographics Workshop on Rendering(2002)
7. Naemura T., Tago J., Harashima H.: Realtime video-based modeling and rendering of 3d scenes. In: IEEE Computer Graphics and Applications 22, 2 (2002), pp.66-73
8. MPEG Document, Survey of Algorithms used for Multi-view Video Coding (MVC). In: ISO/IEC JTC1/SC29/WG11, Doc. N6909, January 2005
9. G. Li, Y. He.: A Novel Multi-View Video Coding Scheme Based on H.264, PCM2003, Singapore, Dec. 2003
10. R.S.Wang, Y.Wang.: Multiview video sequence analysis, compression, and virtual view-point synthesis. In: IEEE Transaction on Circuits and Systems for Video Technology, Vol.10, pp.397-410, April 2000
11. X. Guo, Y. Lu, W. Gao, Q. Huang.: Viewpoint Switching in Multi-view Video Streaming". 2005 IEEE International Symposium on Circuits and Systems(ISCAS 2005), Kobe, Japan, May, 2005
12. X. Guo, Y. Lu, F. Wu, W. Gao, S. Li.: Free viewpoint switching in multi-view video streaming using wyner-ziv video coding. In: SPIE, Visual Communications and Image Processing 2006 (VCIP2006)
13. M.Karczewicz and R.Kurceren.: The SP- and SI-Frames Design for H.264/AVC. In: IEEE Transaction on Circuits and Systems for Video Technology, Vol.13, No.7 July 2003, pp 637-644
14. M. Magnor and B. Girod.: Data compression for light field rendering. In: IEEE Trans. on Circuits and Systems for Video Technology, 10(3):338-343, April 2000
15. M. Magnor and B.Girod.: Hierarchical Coding of Light Fields with Disparity Maps. In: Proc. Int. Conf. Image Processing ICIP-99, Kobe, Japan, pp. 334-338, Oct. 1999
16. A.Isaksen, L.McMillan, and S.J.Gortler.: Dynamically reparameterized light fields. In: Computer Graphics (Proc. SIGGRAPH00), August 2000
17. C.Buehler, M.Bosse, L.McMillan, S.Gortler, and M.Cohen.: Unstructured lumigraph rendering. In: Computer Graphics(Proc. SIGGRAPH01), pp. 425-432, 2001

Complexity Scalability in Motion-Compensated Wavelet-Based Video Coding

T. Clerckx, A. Munteanu, J. Cornelis, and P. Schelkens

Vrije Universiteit Brussel - Interdisciplinary institute for BroadBand Technology,
Dept. of Electronics and Informatics,
Pleinlaan 2, 1050 Brussels, Belgium
tclerckx@etro.vub.ac.be

Abstract. Scalable wavelet-based video codecs based on motion-compensated temporal filtering (MCTF) require complexity scalability to cope with the growing heterogeneity of devices on which video has to be processed. The computational and memory complexity of two spatial-domain (SD) MCTF and in-band (IB) MCTF video codec instantiations are examined in this paper. Comparisons in terms of complexity versus performance are presented for both types of codecs. Some of the trade-offs between complexity and coding performance are analyzed and it is indicated how complexity scalability can be achieved in such video-codecs. Furthermore, a new approach is presented to obtain complexity scalability in IBMCTF video coding, by targeting the complexity of the complete-to-overcomplete discrete wavelet transform at the cost of a limited and controllable penalty on the overall coding performance.

1 Introduction

Real-time delivery of video over best-effort error-prone packet networks requires scalable compression systems in order to (i) meet the users' requirements in terms of quality, resolution and frame-rate, (ii) dynamically adapt the coding rate to the available channel capacity, and (iii) cope with the growing heterogeneity and complexity of devices on which video has to be processed [1].

Wavelet-based architectures based on motion-compensated temporal filtering [2,3,4,5] have proven to be very promising coding systems for scalable video compression. These codecs provide quality, resolution and frame-rate scalability coupled with a coding performance competitive to that of the state-of-the-art H.264 codec [6,7]. Only recently, research has been performed on the complexity of such video coding systems, as in [8] where the encoder's complexity in terms of the number of computations required during motion estimation has been analysed.

In this paper, the complexity of several modules in two instantiations of wavelet-based MCTF video coding architectures are analyzed. For the spatial-domain MCTF [9] codec, both the motion-compensation (MC) and entropy coding module are subject to our research. In addition, in the context of In-Band MCTF (IBMCTF) [10] video coding, the complete-to-overcomplete discrete wavelet-transform (CODWT) [11] is investigated. It is shown that regarding memory-complexity, the CODWT is the most critical component in IBMCTF

video coding. Also, analysis of the results shows how one can reduce complexity and/or obtain complexity scalability by trading off complexity against coding performance.

To overcome the complexity bottleneck of the CODWT, a new approach to obtain complexity scalability in the complete-to-overcomplete discrete wavelet transform is proposed. Experimental results show that fine-grain complexity scalability can be achieved at the cost of a limited and controllable penalty in video coding performance.

The paper is organized as follows. Section 2 overviews the MCTF and the SDMCTF architecture. IBMCTF is described in section 3, together with a description of the shift-variance problem that has to be overcome when applying MCTF in the wavelet-domain. Section 4 reports the setup for the complexity analysis of the different modules under investigation (MC, entropy coding and CODWT) and discusses the obtained results. A mathematical formulation of the CODWT is given in section 5, followed in section 6 by the derivation of a technique that enables complexity scalability in this module. Section 7 evaluates the proposed approach and discusses the obtained results. Final conclusions of our work are drawn in section 8.

2 Spatial-Domain Motion-Compensated Temporal Filtering

MCTF was initially proposed by Ohm and later improved by Choi and Woods [2,3]. It employs an open-loop structure, where frames are temporally filtered along the motion trajectories. MCTF is performed by taking advantage of the

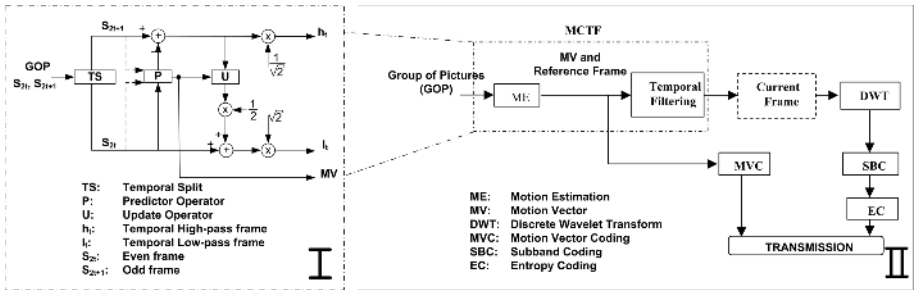


Fig. 1. (a) MCTF scheme, employing the Haar wavelet; (b) Spatial-Domain Motion-Compensated video coding architecture

lifting implementation of the wavelet temporal transform [12,13]. This technique applies the temporal filtering in a sequence of predict and update steps, resulting in a perfectly invertible and computationally efficient process. Figure 1(a) shows the lifting procedure for MCTF decomposition, using the Haar wavelet. The temporal splitting of the input into even and odd frames is followed by motion-compensated prediction, wherein the predict operator P produces an approximation of the odd frames based on the even ones. Within this primal lifting step

the error-frames are determined. Next, within the dual lifting step, the update operator U brings the mismatch information back into the even frames. This step produces a temporal average-frame for each input pair. Finally, the normalization step constructs the low-pass (l) and high-pass (h) temporally-filtered frames respectively. This decomposition process is iterated, using the l -frames at temporal level j as input to produce the l - and h -frames of level $j+1$, until the desired number J of temporal decomposition levels is reached. Spatial-domain motion-compensated temporal filtering (Fig. 1(b)) performs MCTF followed by a 2-D discrete wavelet transform (DWT), yielding a spatio-temporal representation of the input. Subsequently, the spatio-temporally filtered frames are compressed, using an embedded compression scheme and transmitted together with the compressed motion information.

Decoding of the compressed video is obtained by applying the inverse operations.

3 In-Band Motion-Compensated Temporal Filtering

SDMCTF video codecs determine the motion information only at the highest resolution level. This means that decoding at lower resolutions requires downscaling of the motion vectors. The downscaled motion vectors however, do not follow the true motion paths, as they would be obtained by applying motion-estimation at that particular resolution. Consequently, the performance of SDMCT deteriorates when decoding to lower resolutions. Additionally, applying block-based motion-estimation creates artificial boundaries, which, after a spatial wavelet decomposition, results into high-amplitude high-frequency coefficients that are expensive to code. Quantizing these coefficients leads to blocking artefacts that are particularly disturbing when operating at low bit-rates.

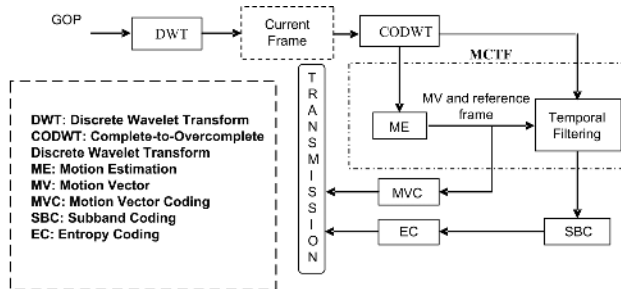


Fig. 2. In-Band Motion-Compensated Temporal Filtering video coding architecture

In-band motion-compensated temporal filtering (IBMCTF) video coding architectures, shown in Fig. 2, do not suffer from these limitations [10]. IBMCTF consists of a DWT front-end followed by an MCTF back-end. IBMCTF applies ME/MC directly in the wavelet domain at every resolution level. In order to apply efficient in-band ME/MC, the wavelet representation of the reference frames

needs to be shift-invariant. The DWT however is only periodically shift-invariant and thus cannot be used for accurate ME/MC. Therefore the overcomplete discrete wavelet representation (ODWT), which is known to be shift-invariant [11], has to be employed in IBMCTF video coding.

Given the input signal, the classical construction of the ODWT is trivial by using for example the “à trous” algorithm [14,15]. However, in wavelet-based coding systems the codec always processes the critically-sampled (complete) DWT subbands. Hence, a complete-to-overcomplete DWT has to take place.

Several solutions have been proposed in literature to calculate the ODWT, given the critically sampled DWT representation. The Low-Band-Shift (LBS) method [16] reconstructs an approximation of the input signal followed by the “à trous” algorithm. This however causes a significant calculation overhead and delay since the input signal has the highest sampling rate. Moreover, LBS is a multi-rate calculation scheme involving a cascade of upsampling and downsampling operations. As a result, even for high-speed high-parallel implementations, the achievable percentage of hardware utilization is low since the filtering of every level has to be pipelined with the production of the results of the previous and the next level.

An alternative that can be used for the calculation of the ODWT is called the Complete-to-Overcomplete Discrete Wavelet Transform (CODWT) [11,17], which calculates directly the ODWT from the critically sampled subbands, using a set of prediction filters. This approach does not require upsampling operations and allows for fast parallel implementations. Moreover, in scenarios that require resolution scalability, the CODWT is computed using a single-rate calculation scheme resulting in a much lower complexity and delay than the LBS-method [11,17].

4 Complexity Analysis of SDMCTF and IBMCTF

4.1 Experimental Setup

In the SDMCTF architecture, both the MC module and the entropy coding module, which is the QuadTree-Limited codec of [18,19] have been examined. For the IBMCTF codec, the access behavior of the CODWT module has been also investigated. Additionally, the complexities of both the IBMCTF and SDMCTF codecs are compared in their integrated form. Given the original frame-rate and resolution, we limit here our analysis to video reconstruction of full or half frame-rate and full or half resolution.

To express the complexity of the modules, accesses to memory has been chosen as a complexity metric. This is justified by the fact that these video-codecs and multi-media applications in general are data dominated. Profiling and instrumentation of the codecs is performed by the PowerEscape tool [20], which is based on the ATOMIUM methodology [21] developed at IMEC.

Note that the reported memory accesses only give a global overview of the complexity behavior. Locality of the data and the size of the memory buffers are also very important figures in complexity analysis, but they were not currently

Table 1. Experimental Setup

	Investigation of QTL and MC	Investigation of SDMCT, IBMCTF and CODWT
Input sequence, resolution	CIF	CIF
Input sequence, frame rate	30 Hz	30 Hz
Temporal levels	4	4
Temporal wavelet filter	HAAR	HAAR
Spatial levels	4	2(IB),4(SD)
Spatial wavelet filter	BIOR2.2	BIOR2.2
Update step	NO	YES
ME accuracy	1/4 pixel	integer
Macroblock sizes	16x16, 8x8	16x16, 8x8

considered in our study. A full description of the parameters used is given in Table 1.

4.2 Results

The access behavior of the QT-L codec as a function of bit-rate is shown in Fig. 3. The results show that for all three test sequences, the access-rate varies quite linearly with the bit-rates. Furthermore, there is not much difference between the graphs for the three sequences, which means that the data-dependency is low. The linearity of the access-rate with the bit-rate is also observed, when decoding to different frame-rates and resolutions, as shown in Fig. 4 for the “Bus” sequence.

From the access-rate perspective, the option providing the largest gain is resolution scaling: accesses performed by QT-L are significantly reduced, especially at high bit-rates, while those caused by the MC should reduce -independently of the bit-rate- to 25% of the amount needed at full resolution. Experiments however report a figure of 33.2%, confirming the expected overhead due to MV scaling and additional interpolations.

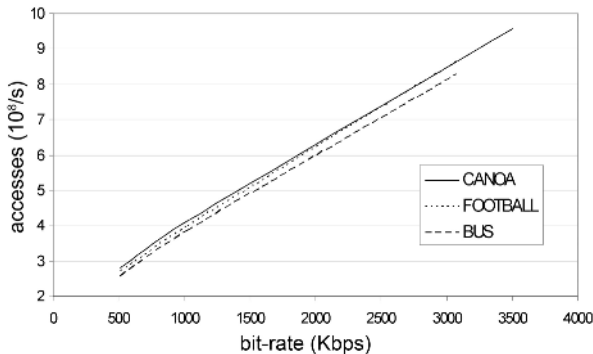


Fig. 3. QT-L: Accesses versus bit rate for different sequences

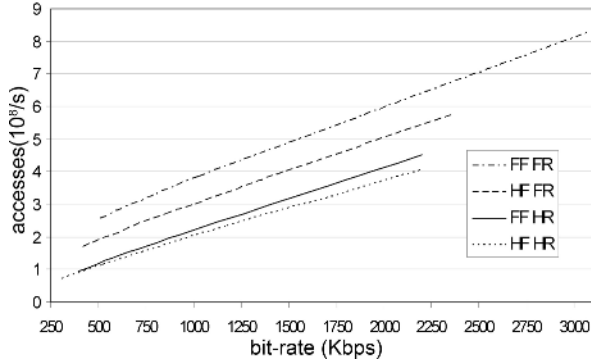


Fig. 4. Accesses versus bit rate for QT-L, in SDMCTF decoding of “Bus” sequence at full/half frame-rate (FF/HF) and full/half resolution (FR/HR)

These results show that the memory access-rate can be decreased in a fine-grain manner by varying the target bit-rate. To avoid reducing the visual quality excessively, the user can switch to a less demanding configuration, decoding lower-resolution versions of the input video or fewer frames, as shown in Fig. 5. Thus, when one resolution or temporal level (or both) are not processed, the decoder switches to an operational point which lays on a curve positioned below the one corresponding to full-resolution and frame-rate decoding. These results show also that the relationship between access-rate and target bit-rate for decoding at different resolutions and frame-rates can be “learned” using appropriate training on large datasets. This way, the decoder can estimate the optimum bit-rate, given a maximum memory access-rate, and a user-specified resolution and frame-rate. Conversely, for bandwidth-limited applications, the decoder can estimate the optimum operational settings in terms of resolution and frame-rate, for a given access-rate and channel-bandwidth.

A comparison between SDMCTF and IBMCTF is shown in Fig. 6. In terms of access-rate, these results clearly show that IBMCTF decoding is about twice as complex as SDMCTF. Both at full frame-rate/full resolution and half frame-rate/half resolution, the difference between the two architectures is almost completely covered by the CODWT module. The small gap between the curve corresponding to IBMCTF decoding at full frame-rate/full resolution for which the CODWT accesses are omitted (FF_FR_inband_no_CODWT) and the curve corresponding to SDMCTF decoding at full frame-rate/full resolution (FF_FR_sd) is justified by the additional operations performed by IBMCTF, as it operates per resolution level (motion compensation, interpolation, addition of the error-frames per subband).

Based on these graphs one can determine which architecture and decoding operational point to choose, given a certain bandwidth or given the processing/memory capabilities of the decoder platform. These results show also that in the IBMCTF architecture, the CODWT module carries half of the complexity. Hence, in order to support a broad set of “complexity profiles”, extensive

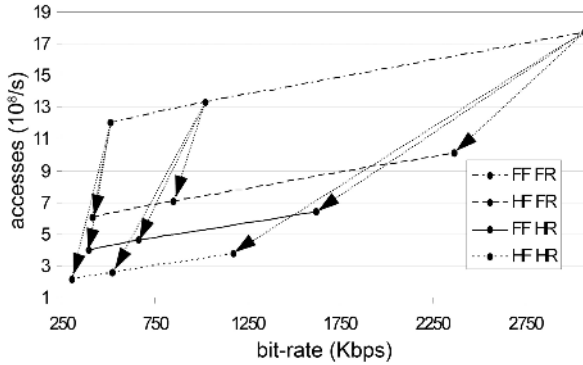


Fig. 5. Access versus bit rate for both MC and QT-L in SDMCTF decoding of “Bus” sequence at full/half frame-rate (FF/HF) and full/half resolution (FR/HR)

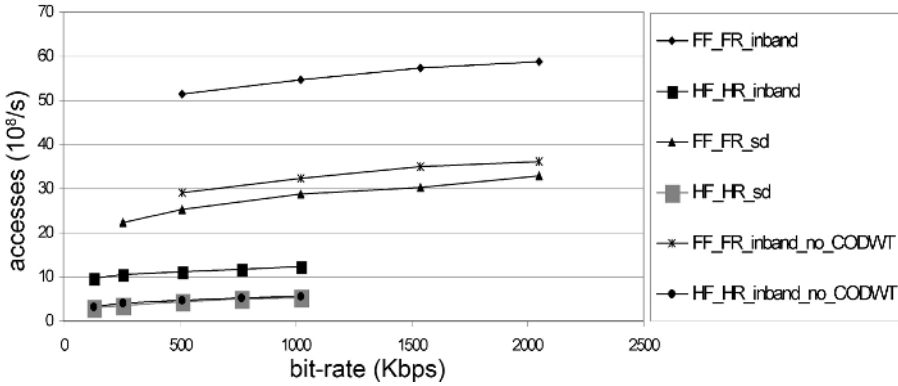


Fig. 6. Decoding of “Bus” sequence at different rates, using IBMCTF (inband) and SDMCTF (sd). The codecs operate at full/half frame-rate (FF/HF) and full/half resolution (FR/HR). The last two graphs show the memory accesses performed by the IBMCTF codec, omitting the accesses made by the CODWT (no_CODWT).

memory optimizations need to be performed, especially on the CODWT module. Additionally, a local approach for the calculation of the CODWT is necessary, as this enables the block-based calculation of the ODWT phases [11,17], leading to significant savings in terms of memory usage. Finally, a complexity-scalable approach is the solution for a progressive complexity reduction of the CODWT. With this respect, the next sections introduce a new method to achieve complexity scalability in the CODWT at the cost of a limited and controllable penalty on the overall coding performance.

5 Mathematical Formulation of the Complete-to-Overcomplete Discrete Wavelet Transform

In an IBMCTF codec, decoding video at a certain resolution level, the ODWT of the reference frames are calculated by applying a set of prediction filters on the critically sampled wavelet-subbands of that level [11]. In the following, a brief description of the CODWT is given for the one-dimensional case, with the extension in two dimensions following the row-column approach [11]. The formulas for the CODWT in the resolution scalable case are given in the Z -domain by:

$$\begin{cases} A_x^k(z) = F_{4p}^{l+1}(z)A_0^k(z) + F_{4p+1}^{l+1}(z)D_0^k(z) \\ D_x^k(z) = F_{4p+2}^{l+1}(z)A_0^k(z) + F_{4p+3}^{l+1}(z)D_0^k(z) \end{cases} \quad (1)$$

where A_x^k and D_x^k denote the low- and high-frequency ODWT-subbands respectively of phase x at resolution level k , F are prediction filters, $l = \lfloor \log_2 x \rfloor$ ($\lfloor a \rfloor$ denotes the integer part of a), and $p = x - 2^l$ [11]. Note that A_0^k, D_0^k correspond to the critically-sampled low- and high-frequency wavelet subbands respectively of level k .

Observations of the prediction filters show that their coefficients tend to have large central values, which decrease rapidly towards the tails [22]. This property can be exploited to reduce the complexity of the CODWT and to obtain complexity-scalability. Using a set of truncated prediction filters reduces the amount of calculations in the filtering process performed by the encoder and/or decoder. The simplest truncation method consists of setting a threshold on the coefficients, which is equivalent to truncating the smallest filter-taps. Next, this method will be referred to as the ‘‘thresholding’’ approach. This technique can be applied in the CODWT calculation performed at the decoder and/or encoder sides, hence enabling complexity scalability at both ends of a video transmission system.

6 Statistical Framework for Prediction Filter Truncation

Truncating the smallest filter-taps does not necessarily imply distortion optimality in a statistical sense, i.e. achieving a minimal average distortion on the calculated ODWT. With this respect, an alternative truncation technique is to identify and truncate those filter taps that yield a minimal distortion in a statistical sense. This approach is described next.

Assume wide-sense stationary wavelet-subbands and zero cross-correlation between different wavelet-subbands. Under these assumptions, based on (1), the expected square error on the low- and high-frequency subbands resulting from truncating the prediction filters can be written in the wavelet-domain as:

$$\begin{aligned} E[(\epsilon_{W_x^k}(n))^2] &= \sum_s \sum_{s'} \epsilon_{4p+j}^{l+1}(s) \epsilon_{4p+j}^{l+1}(s') R_{A_0^k}(s - s') \\ &+ \sum_t \sum_{t'} \epsilon_{4p+j+1}^{l+1}(t) \epsilon_{4p+j+1}^{l+1}(t') R_{D_0^k}(t - t') \end{aligned} \quad (2)$$

with $(W, j) \in \{(A, 0), (D, 2)\}$ and $R_{A_0^k}, R_{D_0^k}$ are the autocorrelation functions of the low- and high-frequency wavelet-subbands respectively. The errors $\epsilon_{4^{p+1}}^{l+1}, \epsilon_{4^{p+j+1}}^{l+1}$, resulting from truncating the prediction filter coefficients follow the same notations as (1) and are either equal to the corresponding prediction filter-taps in case of truncation, either zero. For every combination of truncated coefficients, the expected square error is calculated and the optimal truncated prediction filters are selected as those giving the smallest expected square error. Denote by c the total number of prediction filter coefficients and by p the number of coefficients that are truncated. The total number of possible truncations is then given by $c!(p!(n-p))^{-1}$. As the number of combinations grows factorial, this becomes problematic for a large set of prediction filters. To speed-up these calculations, we use the following empirical property: if we denote by $T_p = \{t_1, t_2, \dots, t_p\}$ the set of coefficient-indices that minimize the expected square error when truncating p filter-taps, then T_{p+1} is given by $T_p \cup \{t_{p+1}\}$, limiting the search space to $pc - (p-1)p/2$ combinations.

7 Experimental Results

First, the truncated prediction filters that are obtained by truncating the smallest coefficients (“thresholding” approach) are employed in the CODWT taking as input a 4CIF-resolution, 3-levels wavelet-transformed image decomposed using a biorthogonal (9-7) wavelet-filter. The distortion in the resulting ODWT expressed in terms of PSNR in the wavelet-domain (W-PSNR), is determined given the reference ODWT, which is obtained by using the non-truncated prediction filters.

The graphs shown in Fig. 7 start at a W-PSNR of 200 for zero truncation points, which is taken as a pseudo-value for lossless calculation of the ODWT. The experiments show that up to 28 prediction filter coefficients can be truncated, while maintaining a W-PSNR above 50 dB for all levels, implying that the losses incurred by truncation on the overall video coding performance are

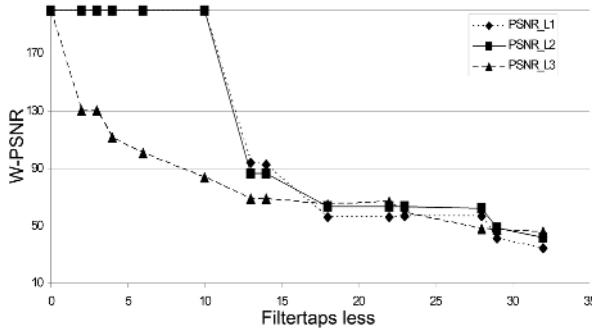


Fig. 7. Wavelet-domain PSNR results for the approximated ODWT for resolution levels 1,2 and 3

minimal. The complexity reduction of the CODWT (see Fig. 8) is measured in terms of accesses to the memory, using the PowerEscape tool [20]. As shown in Fig. 8, the complexity gradually decreases with the number of truncation points, which shows that fine-grain complexity scalability can be achieved, by progressively truncating the prediction filters. Truncating up to 28 coefficients reduces the number of memory accesses of the CODWT with almost 30%, while keeping a W-PSNR above 50dB.

To assess the quality/complexity reduction trade-off globally, the approximated prediction filters are subsequently used in an IBMCTF encoding/decoding scenario of the “Harbour” sequence at a 4CIF-resolution and a frame-rate of 60Hz. Decoding of the sequence at a bit-rate range of 134Kbps-2.39Mbps shows a maximum loss of 1.13dB for different resolutions (QCIF, CIF, 4CIF) and frame-rates (15Hz, 30Hz, 60Hz) when truncating up to 28 filter-taps. The corresponding complexity of the CODWT reduces by almost 30%, which translates to an overall complexity reduction of about 15%, as the CODWT is responsible for almost half the number of memory operations in IBMCTF video coding.

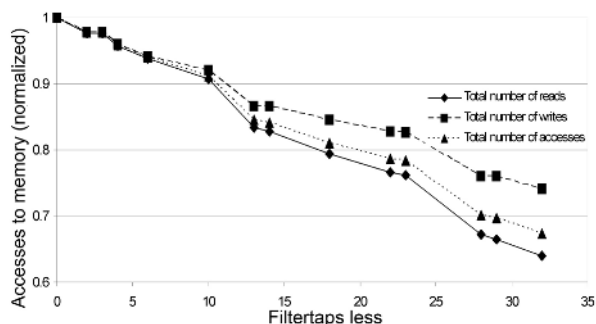


Fig. 8. Normalized number of accesses versus the number of truncated coefficients for the calculation of the full ODWT

In a last experiment, the truncated prediction filters corresponding to a biorthogonal (9-7) wavelet-filter-bank were determined by using the statistical approach and used in the calculation of the CODWT. Here only the distortion needs to be determined, as the complexity reductions in terms of memory accesses are identical to those obtained in the first experiment. The results show that in terms of W-PSNR, the average difference between the statistical and thresholding-based prediction-filter truncation methods is of only 1.2dB. This shows that, although the statistical approach leads to better results, a simple solution such as removing the smallest coefficients results in a quality which is close to the quality obtained when using the statistical distortion-optimal truncation method.

8 Conclusions

Scalable wavelet-based video codecs provide the flexibility needed to adapt the video content to the channel conditions. Moreover, they meet a variety of user preferences in terms of quality, resolution and frame-rate. Apart of this, complexity scalability is of paramount importance in order to deal with the limited and time-varying computational capabilities of the devices on which video has to be processed.

This paper investigates two instances of scalable MCTF wavelet-based video coding (IBMCTF and SDMCTF) architectures. Although both coding systems allow for quality, resolution, temporal and complexity scalability, IBMCTF outperforms SDMCTF when decoding at lower spatio-temporal resolution, but costs almost twice as much in terms of memory accesses compared to the latter. This is mainly caused by the complete-to-overcomplete discrete wavelet transform, which calculates the phases of the ODWT corresponding to the received motion vectors.

Taking into account memory-access rates of the QT-L entropy codec and the MC modules only, it has been shown that the decoder can meet its hardware limitations without requiring transcoding, by an appropriate choice of the quality (bit-rate), frame-rate and/or resolution, thus enabling complexity scalability. To overcome the high memory complexity incurred by the CODWT in the IBMCTF video coding system, a new approach for achieving complexity scalability in this module has been proposed. Although the statistical framework allows for attaining distortion-optimal truncation (in a statistical sense), a simpler technique such as prediction-filter truncation based on thresholding gives results close to the optimal solution. Experimental results demonstrate that the proposed approach achieves fine-grain complexity scalability at the cost of a limited and controllable penalty in the overall video coding performance.

Acknowledgements

This work was supported in part by IWT (PhD bursary T. Clerckx and GBOU Resume Project), BELSPO (IAP Phase V - Mobile Multimedia), and the Fund for Scientific Research - Flanders (post-doctoral fellowships A. Munteanu and P. Schelkens and projects G.0021.03 & G.0053.03).

References

1. Requirements and Applications for Scalable Video Coding, ISO/IEC JTC1/SC29/WG11, Gold Coast, October 2003.
2. Ohm, J.-R.: Three-dimensional subband coding with motion compensation. *IEEE Transactions on Image Processing*, 3 (1994) 559-571.
3. Choi, S.-J., Woods, J. W.: Motion-compensated 3-D subband coding of video. *IEEE Transactions on Image Processing*, 8 (1999) 155-167.

4. Naveen, T., Woods, J. W.: Motion Compensated Multiresolution Transmission of High Definition Video. *IEEE Transactions on Circuits and Systems for Video Technology*, 4 (1994) 29-41.
5. Taubman, D., Zakhor, A.: Multirate 3-D Subband Coding of Video. *IEEE Transactions on Image Processing*, 3 (1994) 572-588.
6. Wiegand, T., Sullivan, G. J., Bjontegaard, G., Ajay, L.: Overview of the H.264/AVC Video Coding Standard. *IEEE Transactions on Circuits and Systems for Video Technology*, 13 (2003) 560-576.
7. Schelkens, P., Andreopoulos, Y., Barbarien, J., Clerckx, T., Verdicchio, F., Munteanu, A., Van der Schaar, M., A comparative study of scalable video coding schemes utilizing wavelet technology, *Proceedings of SPIE Photonics East, Wavelet applications in industrial processing*, Vol. 5266, Providence, (2004) 147-156.
8. D. S. Turaga, M. van der Schaar and Beatrice Pesquet-Popescu, Reduced complexity spatio-temporal scalable motion compensated wavelet video encoding, *IEEE Transactions on Circuits and Systems for Video Technology*, vol. 15, no. 8, Aug. 2005, pp. 982-993.
9. Andreopoulos, I., Barbarien, J., Verdicchio, F., Munteanu, A., van der Schaar, M., Cornelis, J., Schelkens, P., Response to Call for Evidence on Scalable Video Coding, ISO/IEC JTC1/SC29/WG11 (MPEG), Trondheim, Norway, MPEG Report M9911, July 20-25, 2003.
10. Andreopoulos, I., Munteanu, A., Barbarien, J., van der Schaar, M., Cornelis, J., Schelkens, P.: In-band motion compensated temporal filtering. *Signal Processing: Image Communication (special issue on "Subband/Wavelet Interframe Video Coding")*, 19 (2004) 653-673.
11. Andreopoulos, I., Munteanu, A., Van der Auwera, G., Schelkens, P., Cornelis, J.: Complete-to-overcomplete discrete wavelet transforms: theory and applications. *IEEE Transactions on Signal Processing*, 53 (2005) 1398-1412.
12. Flierl, M., Girod, B.: Video Coding with Motion-Compensated Lifted Wavelet Transforms. *Signal Processing: Image Communication*, 19 (2004) 561-575.
13. Pesquet-Popescu, B., Bottreau, V., Three-Dimensional Lifting Schemes For Motion Compensated Video Compression, *Proceedings of IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, Vol., Salt Lake City, (2001) 1793-1796.
14. S. G. Mallat, "A wavelet tour of signal processing.", San Diego: Academic Press, 1998.
15. M. J. Shensa: "The discrete wavelet transform: Wedding the A Trouns and Mallat Algorithms" , *IEEE Transactions on Signal Processing*, vol.40, pp. 2464-2482, 1992.
16. Park, H.-W., Kim, H.-S.: Motion Estimation Using Low-Band-Shift Method for Wavelet-Based Moving-Picture Coding. *IEEE Transactions on Image Processing*, 9 (2000) 577-587.
17. Van der Auwera, G., Munteanu, A., Schelkens, P., Cornelis, J.: Bottom-up motion compensated prediction in the wavelet domain for spatially scalable video coding. *IEE Electronics Letters*, 38 (2002) 1251-1253.
18. Munteanu, A., Wavelet image coding and multiscale edge detection, Department of Electronics and Information Processing (ETRO), Brussel: Vrije Universiteit Brussel, 2003.
19. Schelkens, P., Munteanu, A., Barbarien, J., Galca, M., Giro i Nieto, X., Cornelis, J.: Wavelet Coding of Volumetric Medical Datasets. *IEEE Transactions on Medical Imaging*, 22 (2003) 441-458.
20. PowerEscape, <http://www.powerescape.com/>.

21. ATOMIUM, <http://www.imec.be/design/atomium/>.
22. Andreopoulos, I., Munteanu, A., Van der Auwera, G., Cornelis, J., Schelkens, P.: Single-rate calculation of overcomplete discrete wavelet transforms for scalable coding applications. *Signal Processing*, 85 (2005) 1103-1124.

Spatial Error Concealment with Low Complexity in the H.264 Standard

Donghyung Kim¹, Seungjong Kim², and Jechang Jeong¹

¹ Department of Electrical and Computer Engineering, Hanyang University,
17 Haengdang, Seongdong, Seoul 133-791, Korea
{kimdh, jjeong}@ece.hanyang.ac.kr

² Department of Computer Information, Hanyang Women's College,
17 Haengdang, Seongdong, Seoul 133-793, Korea
jkim@ece.hywoman.ac.kr

Abstract. H.264 adopts new coding tools such as intra-prediction, variable block size, motion estimation with quarter-pixel-accuracy, loop filter, etc. The adoption of these tools enables an H.264-coded bitstream to have more information compared with previous standards. In this paper we proposed an effective spatial error concealment method with low complexity. Among the information included in an H.264-coded bitstream, we use prediction modes of intra-blocks for recovering a damaged block. This is because a prediction direction in each prediction mode is highly correlated to the edge direction. We first estimate the edge direction of a damaged block using prediction modes of intra-blocks adjacent to a damaged block and classify the area inside a damaged block into the edge and the flat area. And then our method recovers pixel values in the edge area using edge-directed interpolation, and recovers pixel values in the flat area using weighted interpolation. Simulation results show the proposed method yields better video quality than conventional approaches.

1 Introduction

Channel noise or congestion often leads to packet loss when video streams are transmitted through noisy channel. As a way to alleviate this problem, error concealment is very useful, since the decoded frame which has lost blocks still includes spatial and temporal redundancy. A spatial error concealment method recovers the lost area using spatially neighboring image data.

Several spatial error concealment algorithms for restoring missing blocks of received video frames have been proposed. Wang et al. proposed the optimization algorithm where the optimal DCT coefficients are estimated by imposing the smoothness constraints between the intensity values of adjacent samples [1]. Lee et al. proposed a spatial error concealment method based on the spatial interpolation filtering and DCT coefficients recovery employing fuzzy logic reasoning [2]. Park et al. proposed a DCT coefficient recovery algorithm for error concealment that has a lower complexity [3]. A block recovery algorithm based on the projection onto convex sets (POCS) was proposed by Sun and Kwok [4] and a fast DCT-based spatial interpolation technique was reported by Alkachouh and Bellanger [5]. A novel error

concealment algorithm, dubbed recovery of image blocks using the method of alternating projections (RIBMAP) was proposed [6].

H.264 adopts new coding tools such as intra-prediction, loop-filter, motion estimation and compensation using variable block size, and so on [7]. The adoption of these tools enables an H.264-coded bitstream to have more information compared with previous standards. Among the information, the prediction mode (pmode) of each intra-block is very useful for spatial error concealment. This is because prediction direction in each pmode is strongly related with the edge direction.

In this paper, using these characteristics of H.264, we present an effective error concealment algorithm with low complexity for intra-frame in the H.264 standard. Using pmodes of intra-blocks adjacent to a lost macroblock, the proposed method estimates the edge direction of a lost macroblock and classifies the damaged area into the edge/flat area in the pre-processing stage. Afterward, the pixel values in the edge area are recovered by edge-directed interpolation and those in the flat area are recovered by weighted interpolation, which is embedded in the reference software of H.264.

2 Intra-mode and Spatial Error Concealment in H.264

2.1 Three Intra-modes of an Intra-block of H.264

To encode a macroblock in an intra-frame by prediction in spatial domain, the H.264 standard uses three different intra-modes: intra16x16, intra8x8 and intra4x4. Each intra-mode predicts the block using the different block sizes. Figure 1 depicts macroblock partitions in case that the macroblock is coded into these three intra-modes, respectively.

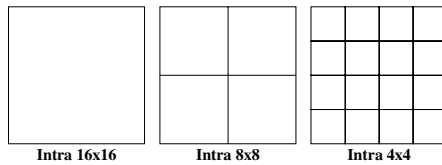


Fig. 1. Block partitions of three intra-modes: intra16x16, intra8x8 and intra4x4

The number of pmodes varies with each intra-mode. Figures 2 and 3 show pmodes for intra16x16 and intra4x4. As shown in these figures, intra16x16 has four pmodes, and intra4x4 has nine pmodes. Because pmodes of intra8x8 are the same as those of intra4x4, intra8x8 also has nine pmodes just like intra4x4.

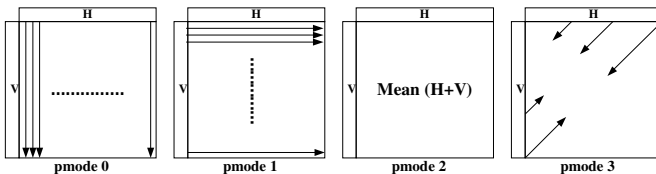


Fig. 2. Four pmodes of intra16x16

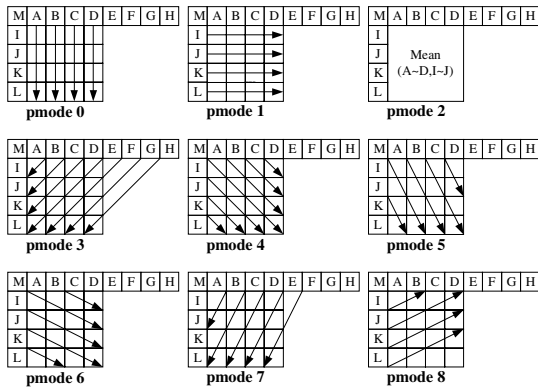


Fig. 3. Nine pmodes of intra4x4

2.2 Spatial Error Concealment in the Reference Software of H.264

For spatial error concealment the reference software (JM 10.1 [8]) of H.264 uses weighted averaging interpolation of four pixel values located at vertically and horizontally neighboring boundaries of a lost macroblock. That is to say, as shown in Fig. 4, a pixel value in a lost macroblock, P_c is replaced with the weighted average of four boundary pixel values located at top, bottom, left and right side. It is formulized in Eq. (1).

$$P_c = (D_1P_0 + D_0P_1 + D_3P_2 + D_2P_3) / (\sum_{i=0}^3 D_i) \tag{1}$$

As shown above, an error concealment method for intra-frames in H.264 takes only the weighted average of vertically and horizontally neighboring boundary pixel values regardless of the edge characteristics of a video frame. This method is relatively effective in the area including no edge (flat area) or vertical/horizontal edges, whereas it causes visual degradations remarkably in the area including the edges of other direction such as diagonal direction.

3 Proposed Algorithm

The proposed algorithm for spatial error concealment in intra-frames during the H.264 decoding process consists of two steps as shown in Fig. 5.

The pre-processing part first decides the dominant prediction mode (DPM) among pmodes around a lost macroblock, and estimates the edge direction of a lost macroblock by the DPM. Also by using the DPM, the damaged area is classified into the edge and the flat area.

After the pre-processing, the adaptive interpolation part restores pixel values with different interpolation methods according to the classified area. The pixel values in the edge area are restored by using an edge-directed interpolation method in advance. And then a weighted interpolation method which is embedded in the reference

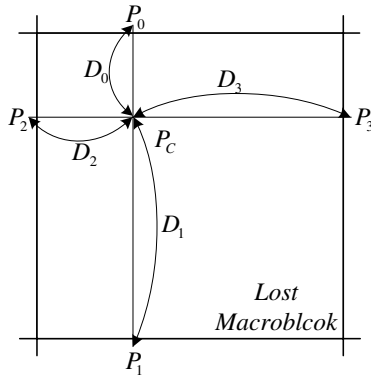


Fig. 4. The spatial error conceal method in the reference software of H.264, where P_c indicates a pixel value in a lost macroblock and P_0 to P_3 are vertically and horizontally neighboring boundary pixel values

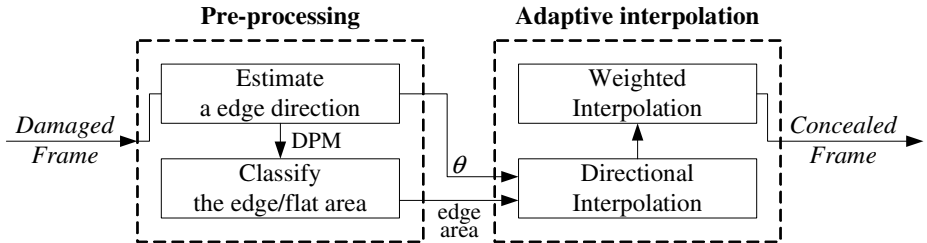


Fig. 5. The architecture of the proposed method for spatial error concealment in intra-frames during the H.264 decoding process

software recovers the pixel values in the flat area using already interpolated pixel values in the edge area as well as the boundary pixel values of a lost macroblock.

3.1 Pre-processing

The proposed method uses pmodes of 4x4 blocks located at top, bottom, left and right side of a lost macroblock. Therefore sixteen pmodes are available. If one neighboring macroblock is coded as intra16x16 or intra8x8, we consider that a pmode of a 4x4 block is repeated. For example, if a macroblock was coded as intra16x16 with using pmode1, we assume that all 4x4 blocks in the macroblock have pmode1. It is because the prediction directions of pmode0 to pmode3 in intra16x16 are identical with those in intra4x4 (see Figs. 2 and 3). Using these pmodes, in the pre-processing stage, we estimate the edge direction of the lost macroblock and classify the edge/flat area.

Estimation of the Edge Direction. Sixteen pmodes located around a lost macroblock is able to be exploited efficiently for the purpose of estimating the edge direction of a lost macroblock, because each pmode represents the edge direction of the corresponding block, as described in Section 2. However, these pmodes do not provide any information

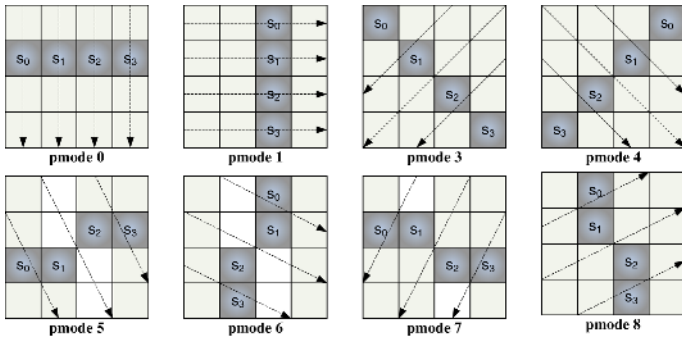


Fig. 6. Four pixel values located at the perpendicular direction of the prediction direction in each pmode for calculating the edge magnitudes of sixteen 4x4 blocks

about the edge magnitude, and just represent the edge directions. Therefore, in order to estimate the edge direction of a lost macroblock, the edge magnitude of each block, as well as pmodes of neighboring blocks, needs to be considered.

The edge magnitudes of sixteen neighboring 4x4 blocks are calculated as follows:

- 1) Deciding four pixels located at the perpendicular direction of the prediction direction in the pmode of each neighboring block.
- 2) Obtaining the difference between maximum and minimum pixel values. It is expressed in Fig. 6 and formulized in Eq. (2). As shown in Eq. (2), for pmode2, we assume that there is no edge, and set the edge magnitude to 0.

$$Edge\ magnitude = \begin{cases} \max(s_i) - \min(s_i) & , pmode \neq pmode2 \\ 0 & , pmode = pmode2 \end{cases} \quad (2)$$

where $i = 0, 1, 2, 3$.

Once the edge magnitudes of sixteen 4x4 blocks around a lost macroblock are obtained, a DPM is decided with a pmode which maximize the sum of the edge magnitude at each pmode, and the prediction direction of a DPM is chosen as the edge direction of a lost macroblock. Thus, the proposed algorithm can consider the eight edge directions identical with prediction direction of pmodes except pmode2. Table 1 depicts the edge direction of a lost macroblock in case that each pmode is chosen as a DPM.

Classification of the Edge/Flat Area. The damaged area can be divided into the edge area where there are edge components and the flat area where there are no edge

Table 1. The relation between eight pmodes except pmode2 and the edge direction (θ) of a lost macroblock

pmode	pmode 0	pmode 1	pmode 3	pmode 4	pmode 5	pmode 6	pmode 7	pmode 8
θ	90°	0°	45°	135°	112.5°	157.5°	67.5°	22.5°

components. Among sixteen 4x4 blocks around a lost macroblock, the block with a pmode decided as a DPM is chosen as a reference block, then the area on the prediction direction of the DPM from the reference blocks is classified as the edge area. In this process, we expand the width of the edge area with some margin because the edges in a natural image smoothly vary differently from ideal edges. Afterward the other area which is not chosen as the edge area is decided as the flat area. Figure 7 depicts the example of the classification of the edge/flat area in case that a pmode4 is chosen as a DPM. If B₁, B₅ and B₁₀ blocks have a pmode4 identical with a DPM, then the area located on the prediction direction of the pmode4 (135°) from the positions of those three blocks with margin considered is chosen as the edge area.

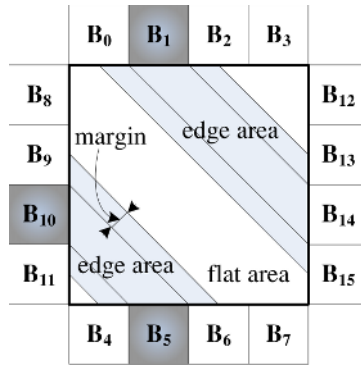


Fig. 7. An example for classification of the edge/flat area when the DPM is pmode4 and B₁, B₅ and B₁₀ blocks have a pmode4 identical with the DPM

3.2 Adaptive Interpolation

After the estimation of the edge direction and the classification between the edge area and the flat area in the pre-processing stage, the lost macroblock is restored by using different interpolation methods for each area. First, for the edge area, pixel values are restored by an edge-directed interpolation method using the estimated edge direction, as depicted in Fig. 8.

In Fig. 8, the pixel value in the edge area, P_C is restored by first-order linear interpolation with two boundary pixel values which are on the straight line from the location of P_C in the direction of the estimated edge direction, as depicted in Eq. (3).

$$P_c = \sum_{i=0}^1 (D_{1-i} \times P_i) / \sum_{i=0}^1 D_i \tag{3}$$

When a DPM is one of pmode5 to pmode8, one of P₀ and P₁ in Fig. 8 is pixel value at a half-pixel position. In this case, the pixel value at a half-pixel position is replaced with the average of two pixel values at neighboring integer positions. Figure 9 depicts the intermediate result of the Foreman sequence, where only pixel values in the edge area are restored. As shown in Fig. 9, most area including edges is decided as the edge area and restored by edge-directed interpolation, whereas the area which represents the flat area is not restored.

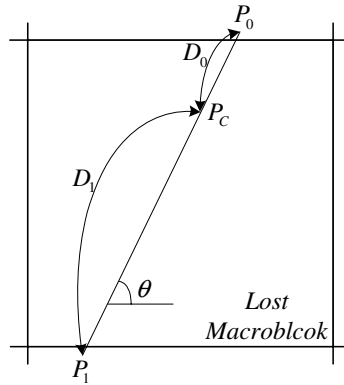


Fig. 8. An edge-directed interpolation method using first-order linear interpolation with two boundary pixel values



Fig. 9. The intermediate result image of the Foreman sequence when only the pixel values in the edge area are restored using the edge-directed interpolation method

3.3 Complexity Analysis

Complexity of the Pre-processing. In the pre-processing, the complexity for estimation of the edge direction depends upon the neighboring pmodes, because there is no operation to calculate the edge magnitude in case of pmode2. In the worst case, that is, if no neighboring 4x4 block is encoded using pmode2, 96(6x16) comparisons and 16(1x16) subtraction per lost macroblock are required for calculating of edge magnitude. The comparison operation is for sorting four values to obtain maximum and minimum pixel values. After the calculation of the edge magnitude of sixteen neighboring 4x4 blocks, 16 additions in order to calculate the sum of edge magnitude on each pmode are required. Lastly, 8 comparisons are needed for selecting a DPM. To selecting of a DPM is equal to estimating the edge direction of a lost macroblock. The classification of the edge/flat area requires only 16 comparisons in order to choose the blocks with the pmode which is identical with the DPM. Consequently,

during the pre-processing, 16 additions, 16 subtractions and 120 comparisons are needed.

Complexity of Adaptive Interpolation. Since the different interpolation methods are used in each area, the complexity of adaptive interpolation varies with the size of each area. The pixel values in the flat area are recovered by using same way embedded in the reference software, as shown in Eq. (1). However, the number of additions in the reference software is less than that of the proposed method in the flat area. It is because the reference software classifies all the damaged area into the flat area (in terms of the proposed algorithm) and fixes the denominator of Eq. (1) as 36, thus saves 3 addition operations. In the proposed method, 6 additions, 4 multiplications and 1 division are required for restoring a pixel value in flat area.

The pixel values in the edge area are recovered by using edge-directed interpolation as shown in Eq. (2). Thus 2 additions, 2 multiplications and 1 division are required for restoring a pixel value in the edge area. Exceptionally, when a DPM is one of pmode5 to pmode8 , 1 addition and 1 shift operation are needed additionally. It is because one of boundary pixels in Eq. (2) is located at a half pixel position.

Comparison of Complexity. As described above, the complexity of the proposed algorithm depends upon pmodes of neighboring blocks, a chosen DPM and the size of each area. Table 2 compares the complexity of the proposed algorithm with one of a method in the reference software of H.264 and Alkachouh's fast DCT-based method when no neighboring block is encoded using pmode2 and the edge area occupies a half of a damaged area.

Table 2. The number of operations for restoring 256 pixel values a lost macroblock when using the reference software, Alkachouh's method and the proposed method in case that no neighboring block is encoded using pmode2 and the half of a lost macroblock is the edge area

Operations	Reference software	Alkachouh's method	Proposed method	
			Pre-processing	Interpolation
Addition	768	17,408	16	1,024 or 1,152
Subtraction	-	-	16	-
Multiplication	1,024	17,408	-	768
Division	256	-	-	256
Comparison	-	512	120	-
Shift	-	-	-	0 or 128

In Table 2, the larger number of the proposed method indicates the case that one of pmode5 to pmode8 is chosen as a DPM, and the number of operations in Alkachouh's method is obtained in case that an interpolation mask (Eq. (20) in [5]) is calculated and saved in advance.

As shown in Table 2, the proposed algorithm estimates the edge direction and classifies the edge/flat area inside a lost macroblock with small amount of operations by using neighboring pmodes known already. Moreover, when the edge area covers more than 50% of the damaged area, the propose algorithm has lower times of multiplication than those of the reference software as well as Alkachouh's.

Considering various edge directions in a video sequence, it is possible for proposed algorithm to alleviate the complexity further.

4 Simulation Results

To evaluate the proposed algorithm, we used a public reference encoder, JVT Model (JM) v.10.1 [7]. Six standard video sequences in CIF (352×288) format were analyzed. These included Akiyo, Coastguard, Container, Foreman, Highway and Silent. We use the first 30 frames of each sequence which were encoded as I-frames only, and compare the simulation results of the proposed algorithm with those of the reference software and Alkachouh's method [5]. In our simulation, the proposed algorithm uses the value 2 as the margin of the edge area.

Figure 10 shows the objective quality for each sequence when 10% of macroblocks are lost at the different bit-rates. As shown in Fig. 10, the proposed algorithm outperforms other two methods in terms of objective quality at all bit-rates for no matter which sequence is considered. Particularly, for Foreman sequence, the

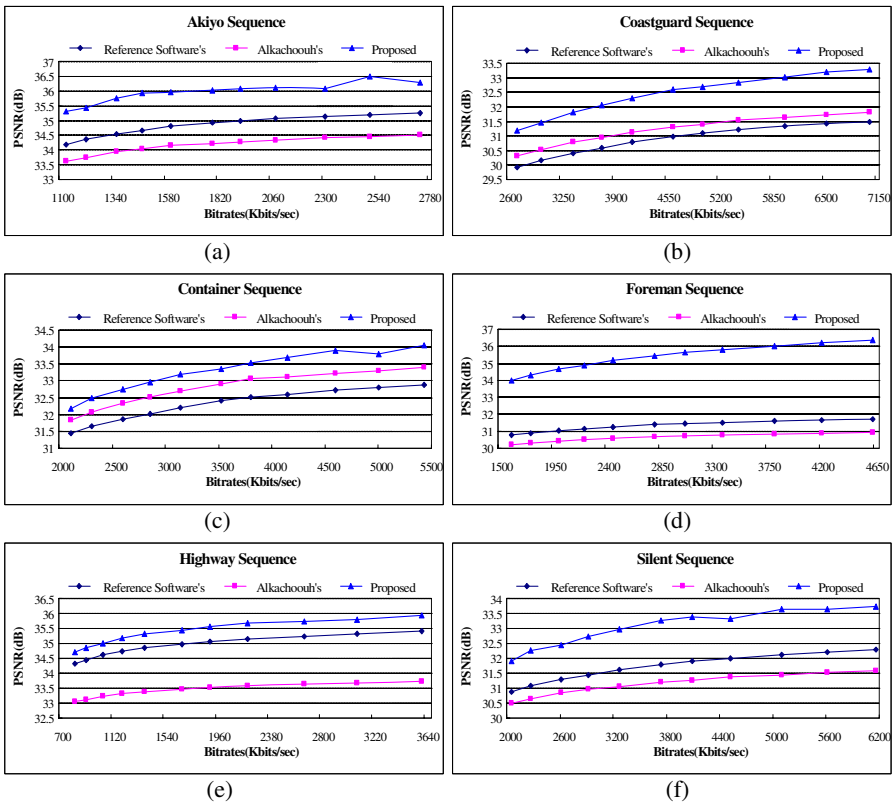


Fig. 10. Comparison of objective qualities at the different bit-rates when using the reference software, Alkachouh's method and the proposed method

proposed algorithm improves the objective quality about 3~5 dB. This is because there are a lot of edge components in the Foreman sequence which can not be considered in those methods.

Figure 11 shows subjective qualities of three methods when 20% of macroblocks are lost in the first frame of the Foreman sequence. As shown in Fig. 11, the proposed algorithm also outperforms compared with the other two methods in terms of subjective qualities.

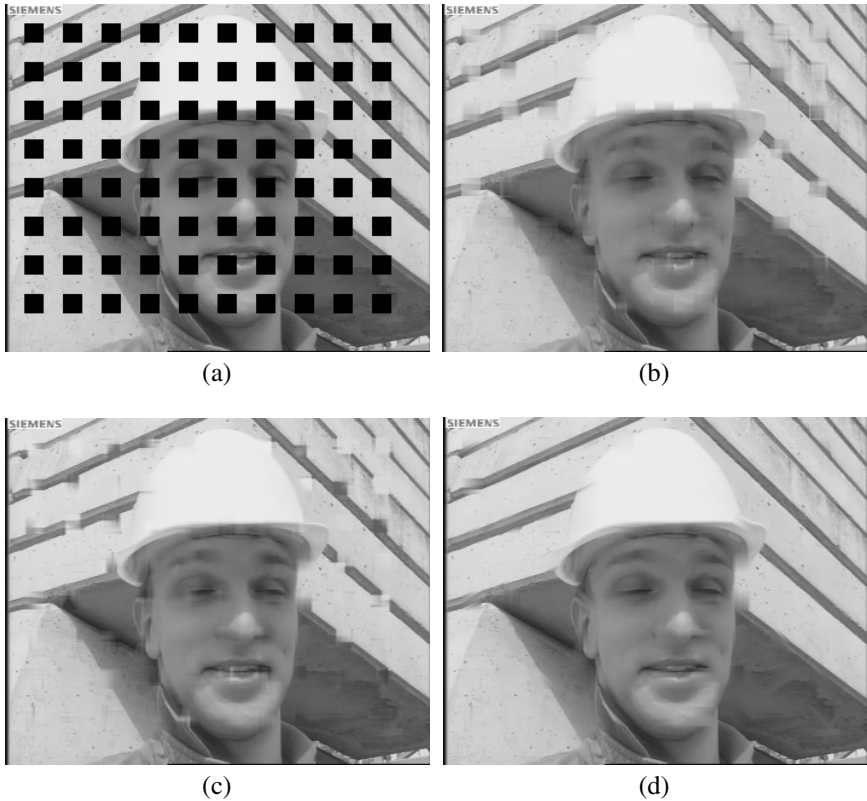


Fig. 11. Comparison of subjective video qualities of the first frame of a foreman sequence (a) a damaged frame (PSNR = 11.26 dB) (b) a recovered frame by the reference software of H.264 (JM10.1 [8]) (PSNR = 29.82 dB) (c) a recovered frame by the Alkachouh's method [5] (PSNR = 28.27 dB) (d) a recovered frame by the proposed method (PSNR = 34.85 dB)

5 Conclusions

The H.264 video coding standard encodes intra-frames using three intra-modes. Each intra-mode has different pmodes, and these pmodes of intra-blocks is included in an H.264-coded bitstream. When the macroblock inside an intra-frame is lost, pmodes included in an H.264 bitstream can be exploited to restore the damaged macroblock

efficiently. This is because each pmode used in encoding an intra-frame is strongly related with the edge direction.

The proposed algorithm conducts the pre-processing which estimates the edge direction of a lost macroblock using pmodes of intra-blocks around the lost macroblock. Also it classifies the area inside the macroblock into either the edge area or the flat area. After that, according to the classified area, different interpolation techniques are applied to restore pixel values. For the edge area, the edge-directed interpolation technique is applied, and for the flat area, the weighted averaging interpolation, which is used in the reference software of H.264, is applied. Besides the proposed algorithm has lower complexity because it makes the utmost use of the properties of the H.264 video coding standard in order to estimate the edge direction and to classify the edge/flat area of the damaged area.

The simulation results show that the proposed algorithm outperforms other two existing techniques for intra-frames in terms of both the subjective and the objective video qualities. Especially it improves the image quality superbly for the sequence in which there are a lot of directions of edge except for a horizontal and a vertical direction.

Acknowledgement

This research was supported by Seoul Future Contents Convergence (SFCC) Cluster established by Seoul Industry-Academy-Research Cooperation Project.

References

1. Wang, Y., Zhu, Q.F., Shaw, L.: Maximally Smooth Image Recovery in Transform Coding. *IEEE Trans. Communication*, Vol. 41. (1993) 1544-1551
2. Lee, X., Zhang, Y.Q., Leon-Garcia, A.: Information Loss Recovery for Block-based Image Coding Techniques-a Fuzzy Logic Approach. *IEEE Trans. Image Processing*, Vol. 4. (1995) 259-273
3. Park, J.W., Kim, J.W., Lee, S.U.: DCT Coefficients Recovery-based Error Concealment Technique and its Application to the MPEG-2 Bit Stream Error. *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 7. (1997) 845-854
4. Sun, H., Kwok, W.: Concealment of Damaged Block Transform Coded Images Using Projections into Convex Sets. *IEEE Trans. Image Processing*, Vol. 4. (1995) 470-477
5. Alkachouh, Z., Bellanger, M.G.: Fast DCT-based Spatial Domain Interpolation of Blocks in Images. *IEEE Trans. Image Processing*, Vol. 9. (2000) 729-732
6. Park, J., Park, D.C., Marks, R.J., El-Sharkawi, M.A.: Recovery of Image Blocks Using the Method of Alternating Projections. *IEEE Trans. Image Processing*, Vol. 14. (2005) 461-474
7. Wiegand, T.: Version 3 of H.264/AVC. Doc. JVT-K051 (2004)
8. JM 10.1: <http://bs.hhi.de/~suehring/tml/download/jm101.zip> (2005)

A Real-Time Content Adaptation Framework for Exploiting ROI Scalability in H.264/AVC

Peter Lambert, Davy De Schrijver, Davy Van Deursen, Wesley De Neve, Yves Dhondt, and Rik Van de Walle

Department of Electronics and Information Systems – Multimedia Lab
Ghent University – IBBT
Gaston Crommenlaan 8 bus 201, B-9050 Ledeborg-Ghent, Belgium
`peter.lambert@ugent.be`

Abstract. In many application scenarios, the use of Regions of Interest (ROIs) within video sequences is a useful concept. It is shown in this paper how Flexible Macroblock Ordering (FMO), defined in H.264/AVC as an error resilience tool, can be used for the coding arbitrary-shaped ROIs. In order to exploit the coding of ROIs in an H.264/AVC bitstream, a description-driven content adaptation framework is introduced that is able to extract the ROIs of a given bitstream.

The results of a series of tests indicate that the ROI extraction process significantly reduces the bit rate of the bitstreams and increases the decoding speed. In case of a fixed camera and a static background, the impact of this reduction on the visual quality of the video sequence is negligible. Regarding the adaptation framework itself, it is shown that in all cases, the framework operates in real time and that it is suited for streaming scenarios by design.

1 Introduction

In many application scenarios, the use of Regions of Interest (ROIs) within video sequences is a useful concept. A ROI typically is a region within the video pane containing visual information that is more interesting than the other parts of the video pane. In the case of multiple ROIs, they can be equally important or they might have different levels of importance. The remaining area is often called the background. Several image or video coding standards (e.g., JPEG2000 [1] or the Fine Granularity Scalability (FGS) Profile of MPEG-4 Visual [2]) have adopted the idea of ROIs and they often provide functionality to code the ROIs at a higher quality level.

The use of ROIs is, for instance, found in surveillance applications. For instance, more and more cameras are developed that capture 360 degrees of video footage with very high resolution pictures. Because it is often impossible to transmit a coded representation of the entire video sequence, one or more ROIs are defined and only a coded version of these smaller areas is transmitted. The position of the ROIs within the picture can mostly be adjusted in real time by an operator. The latter avoids the delays that are introduced by traditional Pan Tilt Zoom (PTZ) cameras.

The currently ongoing standardization efforts of the Joint Video Team regarding Scalable Video Coding (SVC) [3] indicate that there is a clear interest in ROI coding and ROI-based scalability [4,5]. The requirements document of SVC [6] gives some more details about various applications in which ROI coding and ROI-based scalability can be applied, including video surveillance and multi-point video conferencing.

This paper concentrates on the exploitation of ROI coding within the H.264/AVC specification [7]. The H.264/AVC standard does not explicitly define tools for ROI coding, but the authors have shown that the use of *slice groups* (also called Flexible Macroblock Ordering or FMO) enables one to code ROIs into an H.264/AVC bitstream. Notwithstanding the fact that FMO is primarily an error resilience tool, it was illustrated in [8] that it can be the basis for content adaptation. The combination of ROI coding and a description-driven framework for the extraction of ROIs (ROI scalability as content adaptation) is the main topic of this paper. On top of this, it will be shown that the entire content adaptation framework operates in real time and that it is suited for live streaming scenarios. This technique illustrates the possibilities that are offered by the single-layered H.264/AVC specification for ROI-based content adaptation. A similar technique for the exploitation of multi-layered temporal scalability within H.264/AVC is described in [9].

The rest of this paper is organized as follows. Section 2 describes the two main enabling technologies: H.264/AVC FMO and the XML-driven content adaptation framework. In Sect. 3, two methods for ROI extraction are introduced (background slice deletion and placeholder slice insertion). The results of a series of tests regarding the proposed content adaptation framework are given in Sect. 4 and, finally, Sect. 5 concludes this paper.

2 Enabling Technologies

2.1 ROI Coding with H.264/AVC FMO

FMO is a novel tool for error resilience that is introduced in the H.264/AVC specification. By using FMO, it is possible to code the macroblocks of a picture in another order than the default raster scan order (i.e., row per row). One can define up to eight so-called *slice groups* and every macroblock can freely be assigned to one of these slice groups. This assignment results in a MacroBlock Allocation map (MBAmapping), which is coded in a Picture Parameter Set (PPS). In fact, the set of slice groups constitute a *set partition*¹ of the set of macroblocks of a picture. An H.264/AVC encoder will encode one slice group after another and the macroblocks that are part of the slice group in question are coded in raster scan order (*within* that particular slice group). Apart from this, the concept of (traditional) slices remains the same: macroblocks are grouped into slices, the latter being spatially limited to one of the slice groups.

¹ In a strictly mathematical sense.

Because the coding of the entire MBAMap might introduce a considerable amount of overhead, the H.264/AVC standard has specified 6 predefined types of FMO. The MBAMap for these types has a specific pattern that can be coded much more efficiently. FMO type 2, which is used in this paper, indicates that the slice groups are rectangular regions within the video pane, as shown in Fig. 1(a). This type of FMO only requires two numbers to be coded per rectangular slice group. These regions will be considered Regions of Interest. Note that the macroblocks that are left over also constitute a (non-rectangular) slice group. For a thorough overview of H.264/AVC FMO, the reader is referred to [10].

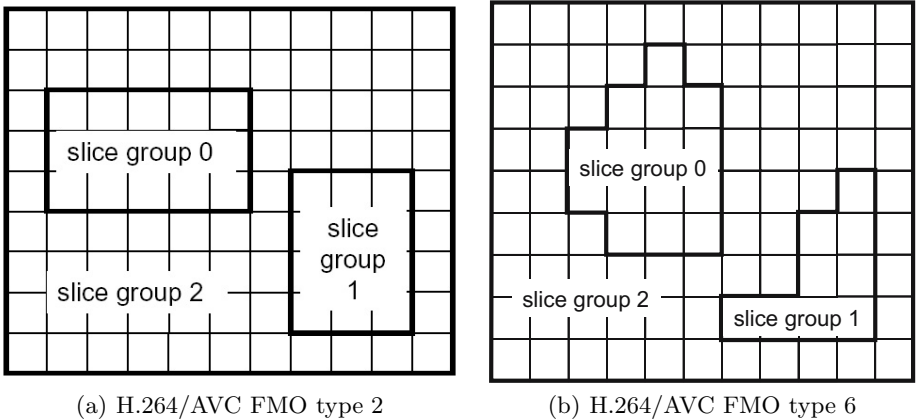


Fig. 1. ROI coding with H.264/AVC FMO

In H.264/AVC, the slice group configuration is coded in a PPS which contains a number of syntax elements that are the same for a certain number of successive pictures (e.g., the entropy coding scheme). Further, every slice contains a reference to the PPS that is into effect. Since the ROI configuration can change in the course of time (e.g., the relative position of a ROI changes, or a ROI appears or disappears), it is required to code a PPS into the bitstream in order to reflect every change of the ROI configuration. In such a PPS, there are four syntax elements that are important in the context of this paper. The number of slice groups is coded by means of `num_slice_groups_minus1` which means that this number denotes the number of ROIs that are present in the bitstream (the ‘background’ is also a slice group). The syntax element `slice_group_map_type` will always be 2 since we only focus on FMO type 2. Every rectangular slice group is defined by the macroblock numbers of its top left and its bottom right macroblock. These two numbers are coded in a PPS by means of the syntax elements `top_left_iGroup` and `bottom_right_iGroup`.

Finally, it should be noted that it is possible to define non-rectangular ROIs in H.264/AVC. Indeed, one can always use FMO type 6 (explicit coding of the MBAMap) to define arbitrary-shaped sets of macroblocks, as depicted in Fig. 1(b). The content adaptation framework, as presented in this paper, is able to

process FMO type 6; the only modification that is needed, is the algorithm that decides if a slice is part of a ROI or not (see Sect. 3.1).

2.2 XML-Driven Content Adaptation Framework

The process of content adaptation based on (pseudo) scalable properties of a bitstream typically requires the removal of certain data chunks, the replacement of certain data blocks, the modification of certain syntax elements, or a combination of these three. One way to accomplish this, is to make use of automatically generated XML descriptions (called Bitstream Syntax Descriptions, or BSDs) that contain high-level information about the bitstreams. For the generation of a BSD, only a limited knowledge is required about the syntax of given bitstream. In stead of performing the adaptations directly on the bitstreams, the generated BSDs can be transformed in such a way that it reflects the desired adaptation. The last step is to automatically generate an adapted bitstream based on the transformed description. The advantage of such an approach is that the adaptation engine itself is truly format agnostic.

The MPEG-21 Bitstream Syntax Description Language (BSDL) framework is an example of a framework that provides the necessary functional blocks that are described above. However, it is described in literature that some parts of this framework have performance issues [11,12], in particular the very high execution times and the monotonically increasing memory consumption of the BintoBSD Parser. As a result, it is (yet) less suited to be deployed in real-life scenarios which require real-time behavior. Another example is the Formal Language for Audio-Visual Object Representation, extended with XML features (XFlavor [13]). The major drawback of the latter is the fact that the generated descriptions are too large because it is required to fully parse the bitstream up to the lowest level (in fact, *all* information of the bitstream is present in its description). In order to combine the strenghts of both BSDL and XFlavor, the authors have developed BFlavor, which is a modification of XFlavor in order to be able to output BSDL-compatible descriptions [14].

BFlavor allows to describe the structure of a media resource in a C++-like manner. It is subsequently possible to automatically create a BS Schema, as well as a code base for a parser that is able to generate a BSD that is compliant with the corresponding BS Schema. This implies that the generated BSDs can be further processed by the upstream tools in a BSDL-based adaptation chain. In Fig. 2, an overview is given of the BSD-oriented content adaptation framework, as employed in this paper. The technology that is used to transform the BSDs is Streaming Transformations for XML (STX, pronounced ‘stacks’) [15]. The internals of the transformation (embodying the actual ROI scalability) are the subject matter of Sect. 3.

3 ROI Extraction

In the context of this paper, every ROI is a slice group (containing one or more slices). Consequently, the extraction of the ROIs comes down to the identification

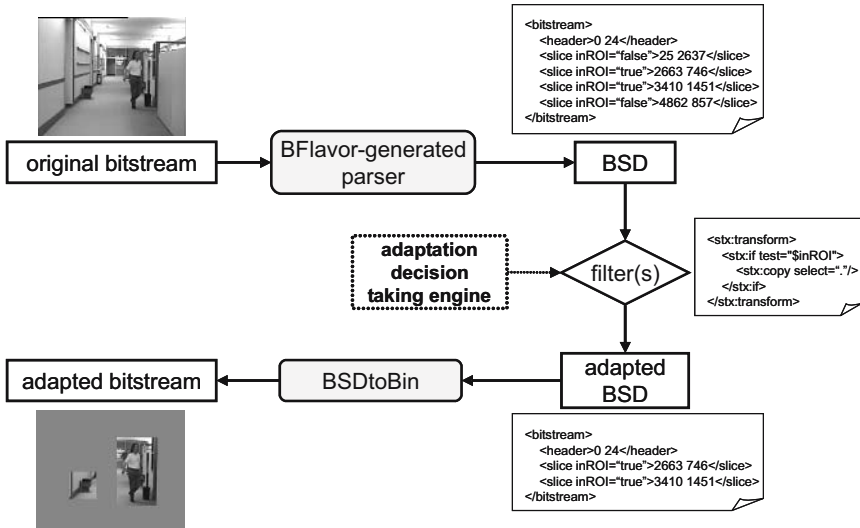


Fig. 2. XML-driven framework for video content adaptation

of those slices that are part of one of the ROIs. Afterwards, the ‘background’ can either be dropped or replaced with other coded data. These two approaches are described in the following two subsections. The bandwidth required to transmit a bitstream that is disposed of its non-ROI parts will be much lower. On top of this, the use of placeholder slices (see Sect. 3.2) will result in a speed-up of the receiving decoder and a decrease in the decoder’s complexity.

3.1 Non-ROI Slice Deletion

For every slice in the coded video sequence, one has to decide whether or not it is part of one of the rectangular slice groups. Based on the syntax element `first_mb_in_slice` (coded in every slice header), this can be done in the following manner. Let R_i be the ROIs and let S be a slice having the macroblock with number FMB_S as its first macroblock (i.e., $FMB_S = \text{first_mb_in_slice}$). Further, let TL_i and BR_i be the macroblock numbers of the top left and bottom right macroblock of ROI R_i . Last, let W be the width of a picture in terms of macroblocks (coded by means of `pic_width_in_mbs_minus1` in a Sequence Parameter Set). Then, S is part of R_i if

$$\begin{aligned}
 & (TL_i \bmod W \leq FMB_S \bmod W) \wedge (FMB_S \bmod W \leq BR_i \bmod W) \\
 & \wedge (TL_i \text{ div } W \leq FMB_S \text{ div } W) \wedge (FMB_S \text{ div } W \leq BR_i \text{ div } W)
 \end{aligned}$$

In this expression, the `div` operator denotes the integer division with truncation and the `mod` operator denotes the traditional modulo operation. Based on a BSD that is generated by BFlavor, this calculation can be done inside a

STX filter. The latter will then discard all parts of the BSD that are related to slices for which the above calculation evaluates to false for all i . It should be noted that I slices are never affected. Based on this transformed BSD, the BSDtoBin Parser can generate the actual adapted bitstream. It is important to note that a bitstream that is generated by this approach will no longer comply with the H.264/AVC standard as the latter requires that *all* slice groups are present in an H.264/AVC bitstream. Despite the fact that only minor modifications of an H.264/AVC decoder are needed for the correct decoding of such a bitstream, this may be considered a disadvantage of the procedure described above.

3.2 Placeholder Slice Insertion

In order to avoid the disadvantages described in the previous subsection, the authors propose the use of placeholder slices. In this approach, coded P and B slices are no longer dropped, but they are replaced by other coded data (again, I slices are not affected). A placeholder slice can be defined as a slice that is identical to the corresponding area of a certain reference picture, or that is reconstructed by relying on a well-defined interpolation process between different reference pictures [16]. Based on the provisions of the H.264/AVC specification, the placeholder slices, as proposed here, are implemented by means of P slices in which all macroblocks are marked as skipped (hereafter called ‘skipped P slices’). This subsection will explain how this substitution can be accomplished in the XML-driven content adaptation framework.

The most straightforward case is replacing coded P slices with skipped P slices. Since the slice header can be kept unchanged, only the slice data are to be substituted. Only two syntax elements are needed to code the slice data for a skipped P slice: `mb_skip_run` to indicate the number of macroblocks that are to be skipped and `rbp_slice_trailing_bits` in order to get byte-aligned in the bitstream. An excerpt of both the original and adapted BSD of a P slice is given in Fig. 3(a) where some simplifications are introduced to improve readability.

In order to replace a coded B slice with a skipped P slice, the substitution process is more complex because of the different nature, and hence header syntax, of P and B slices. The syntax element `slice_type` has to be changed from 1 or 6 (B slice) to 0 (P slice). Next to this, the slice header of a B slice contains a number of syntax elements that cannot appear in a P slice, and they need to be removed. To summarize, the STX filter which adapts the BSDs will remove the following syntax elements (and the syntax elements that are implied by them): `direct_spatial_mv_pred_flag`, `num_ref_idx_l1_active_minus1`, `ref_pic_list_reordering_flag_l1`, `luma_weight_l1_flag` (if applicable), and `chroma_weight_l1_flag` (if applicable).

Regarding the slice data, the same process can be applied as in the case of coded P slices. An example illustrating this scenario is given in Fig. 3(b). In order to save some additional bits, it is possible to change the value of the syntax element `slice_qp_delta` to zero in all cases, as this value has no impact on skipped macroblocks.

```

----- original description -----
<coded_slice_of_a_non_IDR_picture>
  <slice_layer_without_partitioning_rbsp>
    <slice_header>
      <first_mb_in_slice>0</first_mb_in_slice>
      <slice_type>5</slice_type>
      <pic_parameter_set_id>0</pic_p...>
      <frame_num>1</frame_num>
      <!-- ... -->
    </slice_header>
    <slice_data>
      <bit_stuffing>7</bit_stuffing>
      <slice_payload>7875 1177</slice_payload>
    </slice_data>
  </slice_layer_without_partitioning_rbsp>
</coded_slice_of_a_non_IDR_picture>

----- adapted description -----
<coded_slice_of_a_skipped_non_IDR_picture>
  <skipped_slice_layer_without_partitioning_rbsp>
    <slice_header>
      <first_mb_in_slice>0</first_mb_in_slice>
      <slice_type>5</slice_type>
      <pic_parameter_set_id>0</pic_p...>
      <frame_num>1</frame_num>
      <!-- ... -->
    </slice_header>
    <skipped_slice_data>
      <mb_skip_run>108</mb_skip_run>
    </skipped_slice_data>
    <rbsp_trailing_bits>
      <rbsp_stop_one_bit>1</rbsp_stop_one_bit>
      <rbsp_alignment_zero_bit>0</rbsp_a...>
    </rbsp_trailing_bits>
  </skipped_slice_layer_without_partitioning_rbsp>
</coded_slice_of_a_skipped_non_IDR_picture>

```

(a) P slice replaced by a skipped P slice

```

----- original description -----
<coded_slice_of_a_non_IDR_picture>
  <slice_layer_without_partitioning_rbsp>
    <slice_header>
      <first_mb_in_slice>0</first_mb_in_slice>
      <slice_type>6</slice_type>
      <pic_parameter_set_id>1</pic_parameter_set_id>
      <frame_num>2</frame_num>
      <pic_order_cnt_lsb>2</pic_order_cnt_lsb>
      <direct_spatial_mv_pred_flag>1</direct...>
      <num_ref_idx_active_override_flag>1</num...>
      <num_ref_idx_10_active_minus1>1</num...>
      <num_ref_idx_11_active_minus1>0</num...>
      <ref_pic_list_reordering_flag_10>0</ref...>
      <ref_pic_list_reordering_flag_11>0</ref...>
      <slice_qp_delta>2</slice_qp_delta>
    </slice_header>
    <slice_data>
      <bit_stuffing>6</bit_stuffing>
      <slice_payload>9543 851</slice_payload>
    </slice_data>
  </slice_layer_without_partitioning_rbsp>
</coded_slice_of_a_non_IDR_picture>

----- adapted description -----
<coded_slice_of_a_skipped_non_IDR_picture>
  <skipped_slice_layer_without_partitioning_rbsp>
    <slice_header>
      <first_mb_in_slice>0</first_mb_in_slice>
      <slice_type>0</slice_type>
      <pic_parameter_set_id>1</pic_parameter_set_id>
      <frame_num>2</frame_num>
      <pic_order_cnt_lsb>2</pic_order_cnt_lsb>
      <num_ref_idx_active_override_flag>1</num...>
      <num_ref_idx_10_active_minus1>1</num...>
      <ref_pic_list_reordering_flag_10>0</ref...>
      <slice_qp_delta>0</slice_qp_delta>
    </slice_header>
    <skipped_slice_data>
      <mb_skip_run>264</mb_skip_run>
    </skipped_slice_data>
    <rbsp_trailing_bits>
      <rbsp_stop_one_bit>1</rbsp_stop_one_bit>
      <rbsp_alignment_zero_bit>0</rbsp...>
    </rbsp_trailing_bits>
  </skipped_slice_layer_without_partitioning_rbsp>
</coded_slice_of_a_skipped_non_IDR_picture>

```

(b) B slice replaced by a skipped P slice

Fig. 3. XML-driven placeholder slice insertion

4 Results

In order to have some insight in the performance and the consequences of the proposed architecture, a series of tests was set up. The measurements include the impact of the adaptation process on the bitstream and on the receiving decoder. Also an assessment of the performance of the overall adaptation framework is given.

In the experiments, four video sequences were used: Crew (600 pictures with a resolution of 1280 × 720), Hall Monitor, News, and Stefan (the latter three having 300 pictures at CIF resolution). In each sequence, one or more ROIs were manually defined: the moving persons in Hall Monitor and the bag that is left behind by the left person; the heads of the two speakers in News; the tennis player in Stefan; the first two persons of the crew and the rest of the crew as a separate ROI in the Crew sequence. In all sequences, the ROIs are non-static (moving, shrinking, or enlarging) and they may appear or disappear.

These four sequences were encoded with a modified version of the H.264/AVC reference software (JM 9.5) which allows to encode bitstreams with FMO

configurations that vary in the course of time. This encoding was done once conform the Baseline Profile and once conform the Extended Profile (the only difference here being the use of B slices). Other relevant encoding parameters are a GOP length of 16, 2 consecutive B slice coded pictures (if applicable), and a constant Quantization Parameter (QP) of 28. All test runs that are mentioned in this results section were performed on a Pentium IV 2.4 GHz machine with 512MB RAM, running a 2.4.19 Linux kernel. Some properties of the resulting bitstreams are summarized in Table 1. In this table, also the impact of the adaptation process on the bit rate of the bitstreams is given: br stands for the original bit rate, br_p denotes the size of the adapted bitstreams in which placeholder slices were inserted, while br_d denotes the size of the adapted bitstreams of which all background P and B slices are dropped.

Table 1. Bitstream characteristics (sizes in KB)

sequence		# ROIs	# PPSs	# slices	br	br_p	br_d
IP	crew	1-3	48	2020	3856	1379	1376
	hall monitor	1-3	26	924	457	274	272
	news	2	3	904	382	193	190
	stefan	1	31	632	1657	758	756
IBBP	crew	1-3	48	2020	3725	1403	1400
	hall monitor	1-3	26	924	444	277	274
	news	2	3	904	402	193	190
	stefan	1	31	632	1829	819	817

The bitstream sizes clearly indicate that the adaptation process (i.e., ROI extraction) considerably reduces the bit rate required to transmit a bitstream. Both extraction methods (placeholder insertion and background deletion) yield bit rate savings from 38% up to 64%. This reduction has in general a serious impact on the quality of the decoded video sequence. Because the coded background P and B slices are discarded or replaced, a correct picture is only decoded at the beginning of every GOP, resulting in bumpiness of the sequence in which the ROIs are moving smoothly. However, because coded macroblocks inside a ROI can have motion vectors pointing outside the ROIs, ‘incorrect’ decoded data of the background can seep into the ROI which results in erroneous borders of the ROI. This can be avoided by applying so-called *constraint motion estimation* at the encoder so that motion vectors only point to the same slice group the macroblock being predicted belongs to. ROIs that are coded in this way are sometimes called *isolated regions* [17] (this was not used in the tests).

With respect to the (negative) impact of the adaptation process on the received visual quality, there are situations in which this impact is negligible. An example of such a situation is the sequence Hall Monitor in which both the camera and the background are static. The average PSNR-Y of the adapted version is 36.7 dB whereas the unadapted version had an average PSNR-Y of 37.7 dB (or 38.0 dB in case B slices were used). When watching the adapted version, even an expert viewer can hardly notice that the bitstream was subject of an adaptation process. In case of video conferencing or video surveillance applications, this opens up new opportunities. For instance, bitstreams that are coded

with ROIs using H.264/AVC FMO can sustain a rather big decrease in available bandwidth without any noticeable quality loss. This, of course, on condition that the transporting network first ‘drops’ the background packets (e.g., based on priority flags in the network layer). Alternatively, there might be an active network node (implementing the adaptation framework as presented in this paper) which adapts the bitstreams by removing or replacing the coded information of the background.

Because the processing of P-skipped macroblocks requires less operations for a decoder, it is expected that a decoder, receiving an adapted bitstream with placeholder slices, operates faster compared to the case of decoding the original bitstream. Indeed, for the decoding of a P-skipped macroblock, a decoder can rely directly on its decoded picture buffer without performing any other calculations such as motion compensation. Both the original and the adapted bitstreams were decoded five times using the reference decoder (JM 10.2) in order to measure the decoding speed (decoding every bitstream only once could be less reliable). The average decoding speed for each bitstream is given in Table 2.

Table 2. Impact on decoding speed (frames per second)

sequence	original	placeholders	
IP	crew	1.5	2.0
	hall monitor	15.6	16.9
	news	16.7	18.9
	stefan	10.4	14.9
IBBP	crew	1.1	1.3
	hall monitor	13.8	17.0
	news	14.3	17.5
	stefan	9.5	14.4

As can be seen from this table, the decoding speed is positively affected in all cases when placeholder slices are inserted by the adaptation process. The decoding speed in the cases the background was dropped, depends to a great extent on how a receiving decoder copes with non-arriving slices. If a decoder does nothing in case of missing slices, the decoding speed should be higher than the speeds of Table 2. If a decoder performs an error concealment algorithm, the decoding speed will decrease if the applied algorithm is more complex than decoding P-skipped macroblocks (e.g., spatial interpolation techniques).

The last part of this results section is about the performance of the overall adaptation framework. Both the memory consumption and the execution speed are substantial factors for the successful deployment of such an adaptation framework. Therefore, it is important to have an assessment of those factors with respect to the three main components of the adaptation framework as presented in this paper: the generation of BSDs by a BFlavor-generated parser, the transformation of BSDs using STX, and the generation of adapted bitstreams by means of the BSDtoBin Parser. Regarding the memory consumption, it is reported in literature that all components give evidence of a low memory footprint and a constant memory usage [11,12]. As such, the proposed framework satisfies

the memory consumption requirements (i.e., memory consumption is constant in the course of time).

With respect to the execution times of the adaptation framework, every component was executed 11 times both for the placeholder slice insertion method and the background deletion. For all cases, the averages of the last 10 runs are summarized in Table 3. This averaging eliminates possible start-up latencies due to the fact that all components rely on a Java Virtual Machine as their execution environment. In Table 3, the execution speed is given in terms of Network Abstraction Layer Units (NALUs) per second, as a NALU is the atomic parsing unit within the framework. Note that the number of NALUs per picture depends on the slice group configuration. Combining the execution speed of the individual components for both content adaptation methods results in the overall execution speed in terms of frames per second (fps), as denoted in the last two columns of the table.

Table 3. Performance of the overall adaptation framework

sequence	NALUs per second					total fps		
	BFlavor	STX _p	STX _d	BSDtoBin _p	BSDtoBin _d	placeholders	dropping	
IP	crew	1036.3	273.0	308.7	449.2	572.9	43.3	49.9
	hall monitor	2151.1	235.4	272.2	340.3	422.9	46.7	54.9
	news	2371.8	260.6	302.4	344.3	421.5	46.3	54.4
	stefan	1084.5	199.8	264.6	272.2	347.4	49.4	62.4
IBBP	crew	1038.7	221.1	245.5	397.2	497.2	37.1	42.1
	hall monitor	2090.1	200.3	226.0	308.8	385.3	41.0	47.6
	news	2306.4	249.2	284.5	311.7	381.1	43.4	50.5
	stefan	1016.3	155.3	179.1	253.5	325.8	41.8	49.3

It is clear from this table that the proposed framework is capable to perform the content adaptation in real time in all cases (see ‘total fps’). As would be expected, the framework operates slower when performing the placeholder slice insertion because this method requires a more complex transformation in the XML domain. On top of that, the use of B slices also leads to a slow-down in both methods. These two trends can be observed in each component. Notwithstanding the fact that STX is a transformation language that overcomes most performance issues that are encountered when using, for instance, Extensible Stylesheet Language Transformation (XSLT), the transformation in the XML domain still is the slowest component in the framework.

All components of the proposed framework are capable of operating in video streaming scenarios. Indeed, both STX and BSDtoBin are entirely based on SAX events. Although the BFlavor-generated parser currently reads from and writes to a file, it can very easily be modified so that the generated classes use adequate buffers. This streaming capability, and also the performance measurements described above, prove that the proposed framework for the exploitation of ROI scalability within the H.264/AVC specification is suited for real-time video streaming scenarios. This, of course, provided that the identification of the ROIs (motion detection and object tracking) is also done in real time by the encoder and provided that the decoder also operates in real time.

5 Conclusions

In this paper, it was shown how ROI coding can be accomplished within the H.264/AVC video coding specification by making use of Flexible Macroblock Ordering. For the extraction of the ROIs (i.e., exploitation of ROI scalability), a description-driven content adaptation framework was introduced that combines the BFlavor framework for the generation of BSDs, STX for the transformation of these BSDs, and the BSDtoBin Parser of the MPEG-21 BSDL framework for the generation of adapted bitstreams. Two methods for ROI extraction were implemented in this framework by means of a STX filter: removal of the non-ROI parts of a bitstream and the replacement of the coded background with placeholder slices.

Bitstreams that are adapted by this ROI extraction process have a significantly lower bit rate than the original version. While this has in general a profound impact on the quality of the decoded video sequence, this impact is marginal in case of a fixed camera and a static background. This observation may lead to new opportunities in the domain of video surveillance or video conferencing. Next to the decrease in bandwidth, the adaptation process has a positive effect on the receiving decoder: because of the easy processing of placeholder slices, the decoding speed increases.

It was shown that the content adaptation framework, as presented in this paper, operates in real-time. Because each component of the framework is able to function in case of actual streaming video, the framework is suited also suited for live streaming video applications. As such, the framework can be deployed in an active network node, for instance at the edge of two different networks.

Acknowledgements

The research activities as described in this paper were funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), the Belgian Federal Science Policy Office (BFSP), and the European Union.

References

1. Taubman, D., Marcellin, M.: *JPEG2000 : Image Compression Fundamentals, Standards and Practice*. Kluwer Academic Publishers (2002)
2. Li, W.: Overview of fine granularity scalability in MPEG-4 video standard. *IEEE Trans. Circuits Syst. Video Technol.* **11** (2001) 301–317
3. Reichel, J., Schwarz, H., Wien, M.: Joint scalable video model JSVM-4. JVT-Q202, http://ftp3.itu.ch/av-arch/jvt-site/2005_10_Nice/JVT-Q202.zip (2005)
4. Yin, P., Boyce, J., Pandit, P.: FMO and ROI scalability. JVT-Q029, http://ftp3.itu.ch/av-arch/jvt-site/2005_10_Nice/JVT-Q029.doc (2005)

5. Thang, T.C., Kim, D., Bae, T.M., Kang, J.W., Ro, Y.M., Kim, J.G.: Show case of ROI extraction using scalability information SEI message. JVT-Q077, http://ftp3.itu.ch/av-arch/jvt-site/2005_10_Nice/JVT-Q077.doc (2005)
6. ISO/IEC JTC1/SC29/WG11, .: Applications and requirements for scalable video coding. N6880, http://www.chiariglione.org/mpeg/working_documents/mpeg-04/svc/requirements.zip (2005)
7. Wiegand, T., Sullivan, G.J., Bjøntegaard, G., Luthra, A.: Overview of the H.264/AVC video coding standard. *IEEE Trans. Circuits Syst. Video Technol.* **13** (2003) 560–576
8. Dhondt, Y., Lambert, P., Notebaert, S., Van de Walle, R.: Flexible macroblock ordering as a content adaptation tool in H.264/AVC. In: *Proceedings of the SPIE/Optics East conference, Boston* (2005)
9. De Neve, W., Van Deursen, D., De Schrijver, D., De Wolf, K., Van de Walle, R.: Using bitstream structure descriptions for the exploitation of multi-layered temporal scalability in H.264/AVC's base specification. *Lecture Notes in Computer Science, PCM 2005* (2005) 641–652
10. Lambert, P., De Neve, W., Dhondt, Y., Van de Walle, R.: Flexible macroblock ordering in H.264/AVC. *Journal of Visual Communication and Image Representation* **17** (2006) 358–375
11. Devillers, S., Timmerer, C., Heuer, J., Hellwagner, H.: Bitstream syntax description-based adaptation in streaming and constrained environments. *IEEE Trans. Multimedia* **7** (2005) 463–470
12. De Schrijver, D., Poppe, C., Lerouge, S., De Neve, W., Van de Walle, R.: MPEG-21 bitstream syntax descriptions for scalable video codecs. *Multimedia Systems* **11** (2006) 403–421
13. Hong, D., Eleftheriadis, A.: Xflavor: bridging bits and objects in media representation. In: *Proceedings of the International Conference on Multimedia and Expo (ICME), Lausanne, Switzerland* (2002)
14. Van Deursen, D., De Neve, W., De Schrijver, D., Van de Walle, R.: BFlavor: an optimized XML-based framework for multimedia content customization. In: *Proceedings of the Picture Coding Symposium 2006 (PCS 2006)*, accepted for publication (2006)
15. Cimprich, P.: Streaming transformations for XML (STX) version 1.0 working draft. <http://stx.sourceforge.net/documents/spec-stx-20040701.html> (2004)
16. De Neve, W., De Schrijver, D., Van de Walle, D., Lambert, P., Van de Walle, R.: Description-based substitution methods for emulating temporal scalability in state-of-the-art video coding formats. In: *Proc. of WIAMIS, Korea* (2006)
17. Hannuksela, M.M., Wang, Y.K., Gabbouj, M.: Isolated regions in video coding. *IEEE Transactions on Multimedia* **6** (2004) 259–267

Complexity Reduction Algorithm for Intra Mode Selection in H.264/AVC Video Coding

Jongho Kim, Donghyung Kim, and Jechang Jeong

Department of Electrical and Computer Engineering, Hanyang University,
17 Haengdang, Seongdong, Seoul, 133-791, Korea
{angel, kimdh, jjeong}@ece.hanyang.ac.kr

Abstract. The emerging H.264/AVC video coding standard improves coding performance significantly by adopting many advanced techniques. This is achieved at the expense of great increase of encoder complexity. Specifically, the intra prediction using RDO examines all possible combinations of coding modes, which depend on spatial directional correlation with adjacent blocks. There are 9 modes for a 4×4 luma block, and 4 modes for a 16×16 luma block and an 8×8 chroma block, respectively. Therefore the number of mode combinations for each MB is 592. This paper proposes a complexity reduction algorithm using simple directional masks and neighboring modes. The proposed method reduces the number of mode combinations into 132 at the most. Simulation results show the proposed method reduces the encoding time up to 70% with negligible loss of PSNR and bit-rate increase compared with the H.264/AVC exhaustive search.

Keywords: intra mode selection, intra prediction, H.264/AVC, RDO.

1 Introduction

Recent huge requirement for high-performance video codecs has led ISO/IEC MPEG and ITU-T VCEG to develop the video coding standard jointly, known as H.264/AVC [1]. The emerging H.264/AVC standard is incorporated into new applications such as DMB (Digital Multimedia Broadcasting) and DVB-H on account of its good coding performance which is known to be superior to MPEG-4 ASP (Advanced Simple Profile) by about 40% to 50% [2]. The H.264/AVC standard adopts a lot of state-of-the-art techniques to achieve better coding performance: 4×4 block-based integer transform, motion compensation using variable block sizes and multiple references, advanced in-loop deblocking filter, improved entropy coders such as CAVLC (Context Adaptive VLC) and CABAC (Context Adaptive Binary Arithmetic Coding), and enhanced intra-prediction, etc. The RDO (Rate-Distortion Optimization) procedure is conducted in the intra- and inter-prediction of H.264/AVC in order to select the best coding mode among possible mode combinations. The best coding mode from the viewpoint of RDO means that the mode selected among possible mode combinations guarantees the best visual quality under the given bit-rate instead of just minimizing

the bit-rate or maximizing the visual quality. To select the best coding mode based on RDO, the H.264/AVC encoder examines all possible combinations exhaustively. Since transform and entropy coding should be carried out for each coding mode in RDO procedure, it requires very large computational complexity compared to the conventional standards such as MPEG-4 part 2 and H.263, thereby it makes the H.264/AVC standard difficult to apply directly to low complexity devices such as mobile devices. Many algorithms have been proposed to reduce the computational complexity, such as fast motion estimation [3, 4] and fast inter mode selection algorithm [5, 6], etc. Fast motion estimation is a steady-studied subject through various standards and applications. On the other hand, fast mode selection for intra- and inter-prediction in H.264/AVC is a challenging subject. Since there are a lot of mode combinations for each macroblock, fast mode selection of intra- and inter-prediction plays an important role in reducing overall complexity and in applying to various environments. In this proposal, we focus on reducing the complexity of intra mode selection. Intra coding is also carried out in inter-coded frames as well as intra-coded frames, thus fast intra mode selection is valuable for improving the overall coding performance of H.264/AVC.

We propose a complexity reduction algorithm for the H.264/AVC intra-prediction using directional masks for detecting the directional correlation within a block and mode information of adjacent blocks. The proposed directional masks are used for 4×4 luma blocks, since there are many coding modes, on the other hand, for 16×16 luma blocks and 8×8 chroma blocks, we use the mode information of adjacent blocks, since those blocks include a few number of coding modes and the blocks are relatively homogeneous. The directional masks are designed to represent each direction in the H.264/AVC standard. We also address a sampling method in order to reduce computations in SATD (Sum of Absolute Transformed Difference) for 16×16 luma blocks.

The remaining parts of the paper are as follows. We review the intra-prediction scheme of H.264/AVC for 4×4 luma blocks, 16×16 luma blocks, and 8×8 chroma blocks, respectively, and mode selection method based on RDO technique in Section 2. Section 3 presents, in detail, the proposed complexity reduction algorithm in intra mode selection, based on directional masks and mode information of adjacent blocks. Simulation results and conclusions are given in Section 4 and Section 5, respectively.

2 Intra Mode Selection in H.264/AVC

The H.264/AVC intra-prediction exploits the spatial directional correlation with adjacent blocks, and selects the best mode by RDO among a lot of mode combinations. In this section, we review the intra mode decision method for each block type (4×4 luma block, 16×16 luma block, and 8×8 chroma block) of H.264/AVC and address its computational complexity when the RDO procedure is used for the mode selection.

2.1 Intra-prediction in H.264/AVC

One of the advanced features of H.264/AVC compared to the conventional video coding standards is the directional intra-prediction in spatial domain with following considerations: the pixels in spatial domain are more correlated each other than the coefficients in transform domain, and the directional prediction can reflect local characteristics of images better. The intra-prediction, however, requires extremely large computational complexity due to many coding mode combinations in spite of its good coding performance. For the intra-prediction, we use boundary pixels of previously reconstructed adjacent blocks, which are upper, upper-right, and left blocks, and the current block is predicted according to the maximum correlated direction. The H.264/AVC intra-prediction is conducted for all types of blocks such as 4×4 luma blocks, 16×16 luma blocks, and 8×8 chroma blocks. The residual between the current block and its prediction is then transformed, quantized, and entropy coded. For 4×4 luma blocks, which are mainly selected in non-homogeneous areas, there are 9 directional prediction modes, whereas for 16×16 luma blocks, which are selected in relatively homogeneous areas, there are 4 directional prediction modes. In addition, for 8×8 chroma blocks, there are 4 directional prediction modes, and the same mode is applied to two chrominance components (U and V). Note that two types of blocks, i.e., 16×16 luma blocks and 8×8 chroma blocks, have the same directional modes but the order of modes are different from each other.

Fig. 1 shows the 9 intra modes for a 4×4 luma block. In Fig. 1, A to M represent the boundary pixels of previously reconstructed adjacent blocks, which are available at the time of prediction, and the arrows indicate the direction of prediction in each coding mode. DC prediction (*mode 2*) that is not directional mode is carried out using an average of A to L. For *mode 3* to *mode 8*, the pixels of current block are predicted using a weighted average of A to M with the corresponding direction.

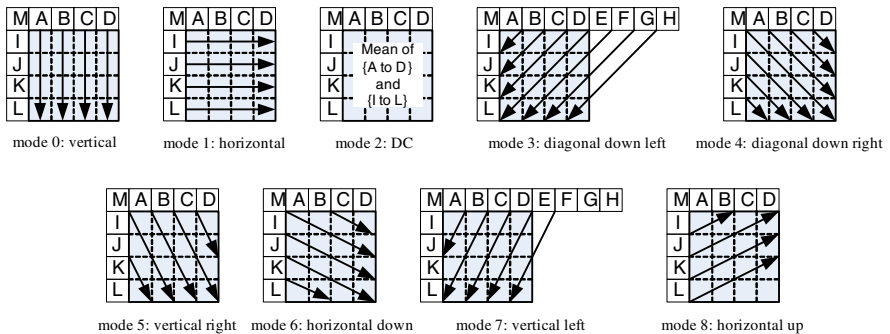


Fig. 1. 9 intra-prediction modes for a 4×4 luma block defined in H.264/AVC

Since 16×16 luma blocks are selected in relatively homogeneous areas mostly, there are fewer prediction modes, i.e., 4 directional modes such as vertical (*mode 0*), horizontal (*mode 1*), DC (*mode 2*), and plane (*mode 3*) prediction. For 8×8 chroma

blocks, there are 4 directional prediction modes, which are very similar to the case of 16×16 luma prediction except the order of modes, such as DC (*mode 0*), horizontal (*mode 1*), vertical (*mode 2*), and plane (*mode 3*) prediction. DC prediction for 8×8 chroma block is carried out with four 4×4 sub-divided blocks using pre-defined adjacent pixels depending on location of the sub-divided block. Both 8×8 chroma blocks of U and V use the same prediction mode. To obtain the best mode among these modes, the H.264/AVC encoder performs the rate-distortion optimization (RDO) technique for each macroblock.

2.2 Selection of the Best Mode Using Rate-Distortion Optimization (RDO)

The RDO procedure for one macroblock in the intra-prediction is as follows [7, 8].

Initialization Set parameters: macroblock quantization parameter QP and Lagrangian multiplier $\lambda_{MODE} = 0.85 \cdot 2^{(QP-12)/3}$ [9].

Step 1 For a 4×4 luma block, select the best mode, which minimizes the *Cost* of (1), among 9 modes.

$$Cost = D + \lambda_{MODE} \cdot R, \quad (1)$$

where D and R denote distortion and bit-rate with given QP, respectively. MODE indicates one of the 9 intra modes of a 4×4 luma block. The distortion is obtained by SSD (Sum of Squared Difference) between the original 4×4 luma block and its reconstructed block, and the bit-rate includes the bits for the mode information and the transformed coefficients for the 4×4 luma block. Repeat this procedure for 16 4×4 luma blocks of a macroblock.

Step 2 For a 16×16 luma block, choose the mode that has the minimum SATD (Sum of Absolute Transformed Difference) among 4 modes as the best mode. In this case, we use Hadamard transform for SATD.

Step 3 For an 8×8 chroma block, select the best mode, which minimizes the $Cost_c$ of (2) among 4 modes.

$$Cost_c = D + \lambda_{MODE} \cdot R, \quad (2)$$

where D is obtained by SSD between two original 8×8 chroma blocks (U and V) and their reconstructed blocks. R, in this case, includes only the bits for the transformed coefficients unlike the 4×4 luma prediction case.

Step 4 Choose the best one as the prediction mode of one macroblock by comparing RD costs for 4×4 mode obtained from Step 1 and 16×16 mode from Step 2.

Considering the RDO procedure for intra mode selection in H.264/AVC, the number of mode combinations in one macroblock is $N_8 \times (16 \times N_4 + N_{16})$, where N_8 , N_4 , and N_{16} represent the number of modes of an 8×8 chroma block, a 4×4 luma block, and a 16×16 luma block, respectively. In other words, to select the best mode for one macroblock in the intra prediction, the H.264/AVC encoder carries out 592 RDO calculations. As a result, the complexity of the encoder increases extremely. We propose, in next section, a complexity reduction algorithm in the H.264/AVC

intra-prediction by reducing the number of RDO calculations for intra mode selection without visual quality degradation.

3 Proposed Complexity Reduction Algorithm

Since the RDO procedure includes time-consuming processes such as transform and entropy coding, the number of RDO computations is a critical point in improvement of the overall encoding speed. To reduce the number of RDO computations, we propose simple and multiplication-free masks, which detect the directional correlation of a block, instead of exhaustive search. Also we use the mode information of adjacent blocks to select the coding mode more accurately. Specifically, in the case of 16×16 luma blocks, we use sampling method to reduce the SATD computations. By using these methods, we reduce the number of RDO computations and encoding time with negligible degradation of visual quality.

3.1 Intra Mode Selection for 4×4 Luma Blocks

We find out two major observations on features of 4×4 luma block as follows: First, the directional correlation of the block is generally consistent with directions of edges. Second, the prediction mode of current block is highly correlated with the modes of adjacent blocks.

From the first observation, we obtain one candidate mode using the proposed directional masks shown in Fig. 2.

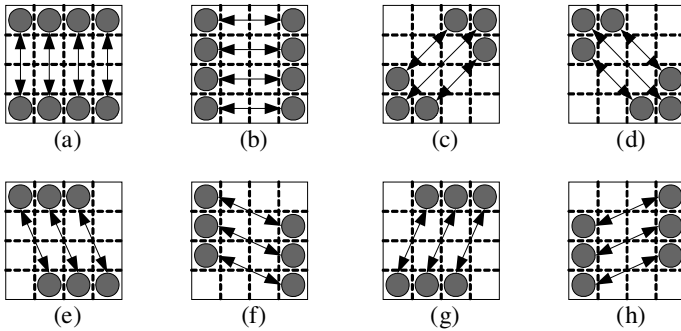


Fig. 2. The proposed directional masks for a 4×4 luma block. (a) vertical, (b) horizontal, (c) diagonal down left, (d) diagonal down right, (e) vertical right, (f) horizontal down, (g) vertical left, (h) horizontal up mask.

In Fig. 2, black dots indicate positions of the pixels to be computed for investing directional correlation in the 4×4 luma block, and arrows represent the directions of correlation associated with the corresponding mask. Since directions of the

H.264/AVC intra-prediction are limited to 8 directions except DC mode, we propose 8 directional masks instead of a precise edge detector such as Sobel operator. We select one candidate mode with the minimum *Diff* using

$$Diff = la - ml + lb - nl + lc - ol + ld - pl, \quad \text{for vertical direction,} \quad (3)$$

$$Diff = la - dl + le - hl + li - ll + lm - pl, \quad \text{for horizontal direction,} \quad (4)$$

$$Diff = lc - il + 2 \cdot ld - ml + lh - nl, \quad \text{for diagonal down left direction,} \quad (5)$$

$$Diff = lb - ll + 2 \cdot la - pl + le - ol, \quad \text{for diagonal down right direction,} \quad (6)$$

$$Diff = la - nl + 2 \cdot lb - ol + lc - pl, \quad \text{for vertical right direction,} \quad (7)$$

$$Diff = la - hl + 2 \cdot le - ll + li - pl, \quad \text{for horizontal down direction,} \quad (8)$$

$$Diff = lb - ml + 2 \cdot lc - nl + ld - ol, \quad \text{for vertical left direction,} \quad (9)$$

$$Diff = le - dl + 2 \cdot li - hl + lm - ll, \quad \text{for horizontal up direction,} \quad (10)$$

where *a* to *p* denote the pixels for investing directional correlation associated with the corresponding mask of Fig. 2. Fig. 3(a) shows the indices for pixel positions used in (3) to (10). *Diff* is used as a criterion for correlation, i.e., the direction with smaller *Diff* is more correlated one.

From the second observation, we obtain additional candidate modes using mode information of adjacent blocks, where one is the upper block with the corresponding mode of *mode_A* and the other is the left block with the corresponding mode of *mode_B*, as shown in Fig. 3(b). We include these additional modes, i.e., *mode_A* and *mode_B*, as candidate modes for RDO procedure, since the directions in the H.264/AVC intra-prediction are defined with the directional relation between current block and boundary pixels of adjacent blocks, instead of direction within the current block only. In this case, one mode when *mode_A* and *mode_B* are same, or two modes when *mode_A* and *mode_B* are different from each other, are included in RDO procedure.

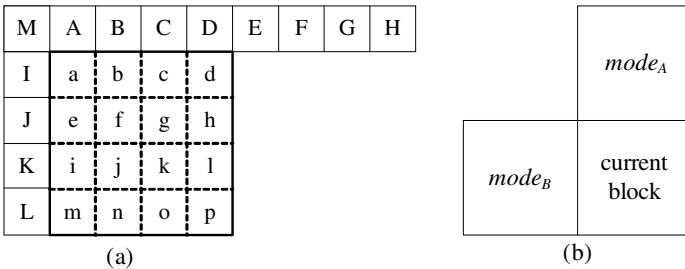


Fig. 3. Pixel indices and modes of adjacent blocks used in the proposed intra mode selection algorithm. (a) indices used in (3) to (10) for a 4x4 luma block, (b) modes of upper and left blocks for additional candidate modes.

To determine whether DC mode is included in RDO procedure or not, we have a sum of difference, denoted by *S* in (11), between an average of current block, denoted by *avg* in (11), and each pixel, denoted by *p_i* in (11).

$$S = \sum_{i=0}^{15} |avg - p_i| \quad (11)$$

where $avg = \left(\sum_{i=0}^{15} p_i + 8 \right) \gg 4$ and p_i is each pixel of current block.

If S is smaller than a threshold, T_I , we carry out RDO for at most 4 candidate modes, i.e., one mode from the proposed masks, at most two modes from adjacent blocks, and DC mode. If S is larger than a threshold, T_I , we carry out RDO for at most 4 candidate modes, i.e., two modes from the proposed masks (with minimum and second minimum $Diff$) and at most two modes from adjacent blocks. The proposed intra mode selection algorithm for a 4×4 luma block is summarized as follows.

Step 1 For a 4×4 luma block, obtain avg and S by (11).

Step 2a If S is larger than a threshold, T_I , carry out RDO procedure for at most 4 candidate modes: two modes with minimum and second minimum $Diff$ by (3) to (10), and at most two modes from adjacent blocks. In this case, DC mode of adjacent blocks is excluded from RDO procedure.

Step 2b If S is smaller than a threshold, T_I , carry out RDO procedure for at most 4 candidate modes: one mode with minimum $Diff$ by (3) to (10), at most two modes from adjacent blocks, and DC mode.

3.2 Intra Mode Selection for 16×16 Luma and 8×8 Chroma Blocks

H.264/AVC carries out the intra-prediction with 16×16 luma blocks when the area to be predicted is relatively homogeneous. The chrominance components of 4:2:0 format are also relatively homogeneous due to down-sampling. Thus, for 16×16 luma blocks and 8×8 chroma blocks, there are only 4 coding modes, different from the case of 4×4 luma blocks, such as horizontal, vertical, DC, and plane mode. For intra mode selection with 16×16 luma blocks and 8×8 chroma blocks, we carry out RDO procedure using the modes of adjacent blocks not using directional masks. Since all adjacent blocks are not 16×16 in this case, we should consider some conditions such as sizes and modes of adjacent blocks when we select the best mode for 16×16 luma blocks. The proposed intra mode selection algorithm for a 16×16 luma block is summarized as follows.

Step 1 Examine sizes of adjacent blocks: if both blocks (upper block and left block) are 16×16 , go to Step 2, otherwise go to Step 4.

Step 2 Examine modes of adjacent blocks: if both modes are same, go to Step 3, otherwise select the best mode for a 16×16 luma block, which results in the minimum SATD between two adjacent modes of $mode_A$ and $mode_B$ shown in Fig. 3(b).

Step 3 If both adjacent modes are DC mode, go to Step 4, otherwise select the best mode for a 16×16 luma block, which results in the minimum SATD between the adjacent mode and DC mode.

Step 4 Let Δ_V be a vertical difference between upper boundary pixels of the current block and boundary pixels of the upper block, and Δ_H be a horizontal difference between left boundary pixels of the current block and boundary pixels of the left block as follows.

$$\Delta_V = \sum_{i=0}^{15} |u_i - q_i|, \quad \Delta_H = \sum_{i=0}^{15} |l_i - r_i|, \quad (12)$$

where u_i and q_i denote boundary pixels of the upper block and upper boundary pixels of the current block, respectively, and l_i and r_i denote boundary pixels of the left block and left boundary pixels of the current block, respectively, as shown in Fig. 4(a). Obtain candidate modes as follows by using two difference values, Δ_V and Δ_H : if $|\Delta_V - \Delta_H|$ is smaller than $2T_2$, candidate modes are DC mode and plane mode; if $(\Delta_V - \Delta_H)$ is larger than T_2 , candidate modes are DC mode and horizontal mode; if $(\Delta_V - \Delta_H)$ is smaller than $-T_2$, candidate modes are DC and vertical mode, where T_2 is a positive value. Finally, select the best mode between each candidate mode by choosing the mode with minimum SATD.

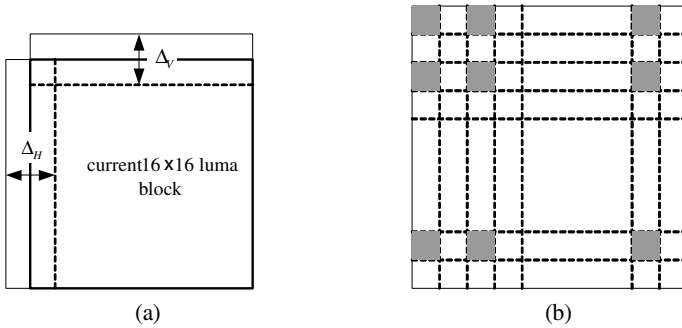


Fig. 4. For intra-prediction with a 16x16 luma block, (a) definition of Δ_V and Δ_H , (b) sampling method to reduce computations

To reduce SATD computation, we propose sampling method as illustrated in Fig. 4(b), where dark shades represent the pixels used for SATD. Since a 16x16 luma block is relatively homogeneous, the mode selection using sampled pixels has almost same results with the mode selection using full pixels. For 8x8 chroma blocks, a similar method to the method for 16x16 luma block is applied except examining whether the sizes of both adjacent blocks are same or not, since all adjacent blocks have the same size in this case.

3.3 Comparison of the Number of RDO Computations

Table 1 summarizes the number of candidate modes for RDO procedure in the proposed method. As it can be seen from Table 1, the proposed algorithm carries out

only 132 RDO computations at the most, which are much less than those of exhaustive search in H.264/AVC video coding, i.e., 592 RDO computations.

Table 1. Comparison of the number of RDO computations

Block type	H.264/AVC method	Proposed method
4×4 luma block	9	at most 4
16×16 luma block	4	2
8×8 chroma block	4	2

4 Simulation Results

In order to evaluate the proposed algorithm, we used JM 8.4 (Joint Model ver. 8.4) provided by JVT (Joint Video Team) under H.264/AVC baseline profile, which contains no B-slice and no CABAC, with RD optimization and Hadamard transform turned on. According to the test conditions specified in [10], we carried out simulations for test sequences of *Akiyo*, *Foreman*, *Carphone*, *Hall Monitor*, *Silent*, *News*, *Container*, and *Coastguard* with QCIF (176×144) resolution. We used various QP of 28, 32, and 40 with IPPP...type and I-only type, respectively. In IPPP...type, the total number of frames is 300 for each sequence, where all frames are inter-coded with one intra-frame for every 100 inter-frames; on the other hand, in I-only type, the total number of frames are 300, where all frames are intra-coded. We compared the results with the case of exhaustive search in terms of the change of average PSNR (Δ PSNR), average data bits (Δ Bit), and average encoding time (Δ Time), respectively, with the machine of Intel Pentium IV processor of 2.8 GHz and 512MB memory.

Table 3 summarizes the simulation results of the proposed algorithm for IPPP type of each sequence. This is because a macroblock can be selected as intra mode in inter-coded frames. In addition, Table 2 shows the results of F. Pan et al.'s method [6] as a reference for comparison. In Table 3, the minus of Δ PSNR and Δ Time means that the encoding time and PSNR are reduced compared with JM, and the thresholds, which include T_1 to select the mode for a 4×4 luma block and T_2 to select the mode for a 16×16 luma block and an 8×8 chroma block, are set to 32 and 8, respectively. It can be seen that the proposed algorithm saves the encoding time up to about 35% with negligible loss in PSNR and increment in bits. By comparing Tables 2 and 3, we can see that the proposed algorithm is superior to F. Pan et al.'s method. This is because F. Pan et al.'s method, to reduce the RDO computation, performs the quite complex pre-processing, and does not use the mode information of adjacent blocks.

We also show the simulation results for I-only type sequences in Table 4. The thresholds, in this case, are set to the same values as the case of IPPP type sequences. It can be seen that the proposed algorithm saves the encoding time up to about 70%, since all frames are intra-coded. On the other hand, the drop in PSNR and the increase of bit-rate is somewhat larger than the IPPP case, since all frames are selected as intra

mode in I-only case, whereas the drop in PSNR and the increase of bit-rate is negligible in IPPP case, since the number of macroblocks selected as intra mode over a sequence is limited to small. However, these results can be acceptable because the drop in PSNR and the increase of bit-rate of I-only case are considerably small with respect to saving the encoding time, and I-only case is regarded as an extreme case. By comparing Tables 2 and 4, we can see that the proposed algorithm is superior to F. Pan et al.'s method for the same reasons of the IPPP case.

Fig. 5 and Fig. 6 show the R-D curve, which includes the results of JM, F. Pan et al., and the proposed method, for IPPP type and I-only type of *News* sequence, respectively. One can see that the proposed algorithm is superior to JM and F. Pan et al.'s method as similar to Tables 3 and 4.

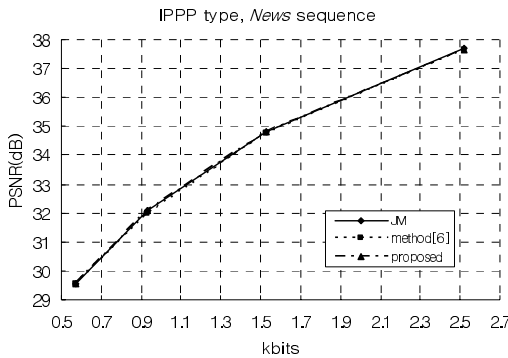


Fig. 5. R-D curve for *News* sequence, QCIF, IPPP type

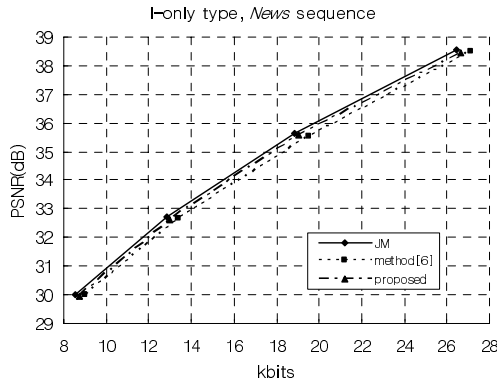


Fig. 6. R-D curve for *News* sequence, QCIF, I-only type

Table 2. Results of F. Pan et al.'s method for comparison

Sequence (QCIF)	IPPP type			I-only type		
	Δ Time (%)	Δ PSNR (dB)	Δ Bit (%)	Δ Time (%)	Δ PSNR (dB)	Δ Bit (%)
Akiyo	-22.72	-0.053	1.17	-64.32	-0.210	3.21
Foreman	-21.80	-0.077	1.54	-65.38	-0.285	4.44
Carphone	-20.51	-0.082	1.80	-65.93	-0.276	3.91
Hall Monitor	-23.38	-0.065	1.23	-66.51	-0.252	3.73
Silent	-21.94	-0.033	0.86	-65.17	-0.183	3.54
News	-23.11	-0.067	1.23	-55.34	-0.294	3.90
Container	-20.78	-0.081	1.80	-56.36	-0.234	3.70
Coastguard	-21.20	-0.017	0.50	-55.03	-0.106	2.36

Table 3. Simulation results for IPPP type sequences

Sequence (QCIF)	QP = 28			QP = 32			QP = 40		
	Δ Time (%)	Δ PSNR (dB)	Δ Bit (%)	Δ Time (%)	Δ PSNR (dB)	Δ Bit (%)	Δ Time (%)	Δ PSNR (dB)	Δ Bit (%)
Akiyo	-29.65	-0.013	0.16	-32.94	-0.008	0.21	-35.56	-0.002	0.31
Foreman	-30.61	-0.016	0.22	-34.37	-0.010	0.28	-36.17	-0.001	0.34
Carphone	-30.87	-0.018	0.17	-34.52	-0.013	0.32	-35.91	-0.009	0.33
Hall Monitor	-33.12	-0.018	0.20	-35.31	-0.014	0.26	-38.35	-0.005	0.37
Silent	-32.05	-0.014	0.18	-34.23	-0.008	0.25	-36.48	-0.004	0.30
News	-33.54	-0.017	0.19	-35.81	-0.011	0.25	-37.63	-0.003	0.30
Container	-29.93	-0.018	0.16	-33.42	-0.011	0.22	-35.41	-0.003	0.27
Coastguard	-30.05	-0.010	0.12	-33.62	-0.007	0.18	-35.42	-0.001	0.24

Table 4. Simulation results for I-only type sequences

Sequence (QCIF)	QP = 28			QP = 32			QP = 40		
	Δ Time (%)	Δ PSNR (dB)	Δ Bit (%)	Δ Time (%)	Δ PSNR (dB)	Δ Bit (%)	Δ Time (%)	Δ PSNR (dB)	Δ Bit (%)
Akiyo	-62.34	-0.10	0.47	-67.16	-0.12	0.93	-68.85	-0.07	1.62
Foreman	-62.91	-0.08	0.15	-68.11	-0.07	1.06	-70.32	-0.03	1.79
Carphone	-65.28	-0.16	0.92	-67.84	-0.13	1.52	-69.13	-0.10	1.73
Hall Monitor	-66.12	-0.14	0.36	-69.63	-0.12	1.92	-71.42	-0.10	2.94
Silent	-63.35	-0.12	0.51	-67.35	-0.08	1.35	-60.04	-0.05	2.63
News	-61.38	-0.11	0.86	-66.56	-0.10	1.26	-69.28	-0.06	1.85
Container	-61.93	-0.09	0.90	-67.07	-0.08	1.07	-69.21	-0.05	1.71
Coastguard	-60.82	-0.08	0.73	-65.81	-0.06	0.82	-68.23	-0.03	1.57

5 Conclusions

This paper has presented a complexity reduction algorithm for intra mode selection in H.264/AVC based on directional masks and mode information of adjacent blocks.

The proposed directional masks have simple structures, which require no multiplications. The simulation results show that the proposed algorithm reduces the number of mode combinations and computational complexity for RDO with negligible loss of PSNR and bit-rate increment. The proposed algorithm can be applied to the H.264/AVC video encoder with low computational capability.

Acknowledgments. This work was supported by the Ministry of Commerce, Industry and Energy with the project of development of personal next generation TV terminal.

References

1. ITU-T Rec. H.264 | ISO/IEC 14496-10: Information Technology – Coding of Audio-visual Objects, Part 10: Advanced Video Coding (2002)
2. Wiegand, T., Sullivan, G., Bjontegaar, G., Luthra, A.: Overview of the H.264/AVC Video Coding Standard. *IEEE Trans. Circuits and Syst. for Video Technol.*, Vol. 13. (2003) 560-576
3. Chen, Z., Zhou, P., He, Y.: Fast Integer Pel and Fractional Pel Motion Estimation for JVT. Doc. JVT-F017 (2002)
4. Hsieh, B., Huang, Y., Wang, T., Chien, S., Chen, L.: Fast Motion Estimation for H.264/MPEG-4 AVC by Using Multiple Reference Frame Skipping Criteria. VCIP 2003, Proceedings of SPIE, Vol. 5150. (2003) 1551-1560
5. Lim, K., Wu, S., Wu, D., Rahardja, S., Lin, X., Pan, F., Li, Z.: Fast Inter Mode Decision. Doc. JVT-I020 (2003)
6. Pan, F., Lin, X., Rahardja, S., Lim, K., Li, Z., Wu, D., Wu, S.: Fast Mode Decision Algorithm for Intraprediction in H.264/AVC Video Coding. *IEEE Trans. Circuits and Syst. for Video Technol.*, Vol. 15. (2006) 813-822
7. Kim, C., Shih, H., Kuo, C.: Multistage Mode Decision for Intra Prediction in H.264 Codec. VCIP 2004, Proceedings of SPIE, Vol. 5308 (2004) 355-363
8. Lim, K., Sullivan, G., Wiegand, T.: Text Description of Joint Model Reference Encoding Methods and Decoding Concealment Methods. Doc. JVT-N046 (2005)
9. Stockhammer, T., Kontopodis, D., Wiegand, T.: Rate-distortion Optimization for JVT/H.26L Video Coding in Packet Loss Environment. *Int. Packet Video Workshop* (2002)
10. Sullivan, G.: Recommended Simulation Common Conditions for H.26L Coding Efficiency Experiments on Low Resolution Progressive Scan Source Material. Doc. VCEG-N81 (2001)

Simple and Effective Filter to Remove Corner Outlier Artifacts in Highly Compressed Video

Jongho Kim, Donghyung Kim, and Jechang Jeong

Department of Electrical and Computer Engineering, Hanyang University,
17 Haengdang, Seongdong, Seoul, 133-791, Korea
{angel, kimdh, jjeong}@ece.hanyang.ac.kr

Abstract. We propose a detection method of corner outlier artifacts and a simple and effective filter in order to remove the artifacts in highly compressed video. We detect the corner outlier artifacts based on the direction of edges going through a block corner and the properties of blocks around the edges. Based on the detection results, we remove the stair-shaped discontinuities, i.e., corner outlier artifacts, using the neighboring pixels of the corner outlier artifact in the spatial domain. Simulation results show that the proposed method improves, particularly in combination with a deblocking filter, both objective performance and subjective visual quality.

Keywords: corner outlier artifact, MPEG-4 video, deblocking filter, highly compressed video.

1 Introduction

Most video coding standards, which have a hybrid structure, adopt block-based motion compensated prediction and transform. Consequently, the block-based processing generates undesired artifacts such as blocking artifacts, ringing noise, and corner outlier artifacts, particularly when the video is highly compressed. Blocking artifacts are grid noise along block boundaries in relatively flat areas; ringing noise is the Gibb's phenomenon owing to truncation of high-frequency coefficients by quantization; corner outlier artifacts are a special case of blocking artifacts at the cross-point of a block corner and a diagonal edge. To reduce the blocking artifacts and the ringing noise, a number of studies have been carried out in the spatial domain [1, 2, 5] and transform domain [3, 4], respectively. However, the corner outlier artifacts are still visible in some video sequences, since a deblocking filter is not applied to the areas including a large difference at a block boundary in order to avoid undesired blurring [1]. Hardly any studies have been carried out on removing the corner outlier artifacts although the artifacts degrade visual quality considerably because the corner outlier artifacts appear only in limited areas and the peak signal-to-noise ratio (PSNR) improvement is somewhat small. Therefore, we propose a detection method of the corner outlier artifacts and a simple and effective filter to remove the artifacts in highly compressed video. The proposed method can be used along with various deblocking [1-4] and deringing filters [5] for more improvement of visual quality.

The remaining parts of the paper are as follows. We present a detection method of the corner outlier artifacts based on the pre-defined patterns, i.e., the direction of an edge going through a block corner, in Section 2. Section 3 describes, in detail, the proposed filter to remove the corner outlier artifacts. Simulation results and conclusions are given in Section 4 and Section 5, respectively.

2 Detection of Corner Outlier Artifacts

Consider the original and its reconstructed frames illustrated in Fig. 1, where a diagonal edge goes through a block corner. The edge occupies large areas in blocks **B** and **C**, whereas it occupies very small areas (d_0 in Fig. 1) in block **D**. If **D** is flat except d_0 , the AC coefficients of DCT of block **D** are mainly related to d_0 , and their values are small. Since most small AC coefficients are truncated by quantization in very low bit-rate coding, the area of d_0 , which represents the edge in block **D**, cannot be reconstructed as shown in Fig. 1(b). As a result, the visually annoying stair-shaped artifact is produced around the block corner. Such artifacts are called corner outlier artifacts.

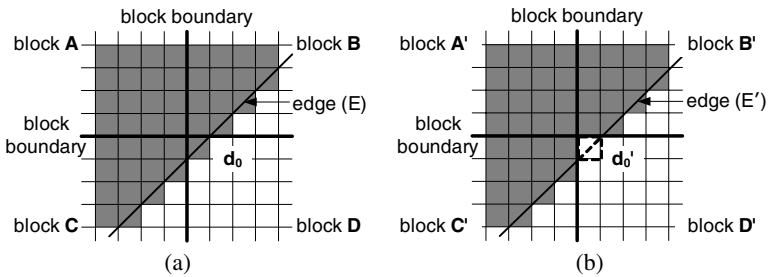


Fig. 1. Conceptual illustration of a corner outlier artifact, (a) an edge going through a block corner in an original frame, (b) an corner outlier artifact located in d_0' in a reconstructed frame

Based on two major observations, we propose a simple detection method for corner outlier artifacts. First, there is a large difference between boundary values of the block including the corner outlier artifact and the other three blocks around the cross-point. For example, the difference between d_0' and its upper pixel or between d_0' and its left pixel is large as shown in Fig. 1(b). Second, corner outlier artifacts are more noticeable in flat areas, that is, each block around the cross-point is relatively flat. We examine the characteristics of the blocks in terms of the observations around every cross-point to detect the corner outlier artifacts appropriately. In addition, since the corner outlier artifacts are prominent when a diagonal edge occupies block areas unequally around a cross-point, we deal with four types, as depicted in Fig. 2, based on the edge direction. We detect the actual corner outlier artifact, which satisfies the proposed detection conditions among each detection type. Dealing with the pre-defined detection types instead of detecting an edge precisely has an advantage in reduction of computational complexity.

To detect corner outlier artifacts based on the first observation, we obtain the average values of the four pixels around the cross-point, which are represented as the shaded areas in Fig. 3, by

$$A_{avg} = \frac{1}{4} \sum_{i=1}^4 a_i, \quad B_{avg} = \frac{1}{4} \sum_{i=1}^4 b_i, \quad C_{avg} = \frac{1}{4} \sum_{i=1}^4 c_i, \quad D_{avg} = \frac{1}{4} \sum_{i=1}^4 d_i \quad (1)$$

where each capital and small letter denotes blocks and pixels, respectively, that is, A_{avg} is an average from a_1 to a_4 of block **A**, B_{avg} is an average from b_1 to b_4 of **B**, and so on.

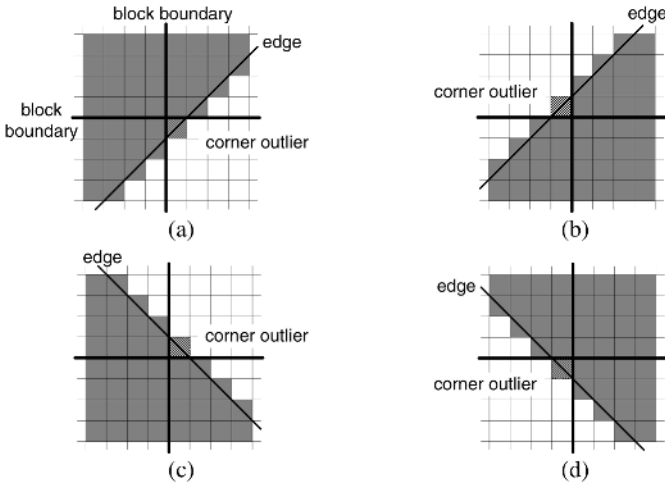


Fig. 2. Detection types based on the edge direction, (a) Lower 45° direction, (b) Upper 45° direction, (c) Upper 135° direction, (d) Lower 135° direction

	block boundary						
block A							block B
	a ₇	a ₆	a ₅	b ₅	b ₆	b ₇	
	a ₈	a ₄	a ₃	b ₃	b ₄	b ₈	
	a ₉	a ₂	a ₁	b ₁	b ₂	b ₉	
block boundary	c ₉	c ₂	c ₁	d ₁	d ₂	d ₉	
	c ₈	c ₄	c ₃	d ₃	d ₄	d ₈	
	c ₇	c ₆	c ₅	d ₅	d ₆	d ₇	
block C							block D

Fig. 3. Indices of pixels for detecting and filtering corner outlier artifacts

Then, to determine the detection type among the four cases, we examine the differences between the average values of the corner-outlier-artifact candidate block

and its neighbors, using equations listed in Table 1. By using the average values around the cross-point instead of each pixel value, we can reduce detection errors in complex areas.

Table 1. Detection criterion according to the detection types

Block including the artifact	Type	Detection criterion
A	A vs. { B, C, D }	$ A_{avg} - B_{avg} > 2QP$ and $ A_{avg} - C_{avg} > 2QP$
B	B vs. { A, C, D }	$ B_{avg} - A_{avg} > 2QP$ and $ B_{avg} - D_{avg} > 2QP$
C	C vs. { A, B, D }	$ C_{avg} - A_{avg} > 2QP$ and $ C_{avg} - D_{avg} > 2QP$
D	D vs. { A, B, C }	$ D_{avg} - B_{avg} > 2QP$ and $ D_{avg} - C_{avg} > 2QP$

In Table 1, QP denotes the quantization parameter. For example, when block **A** has a corner outlier artifact, the differences between A_{avg} and B_{avg} and between A_{avg} and C_{avg} is large by the first observation, thereby we consider the block **A** has a corner outlier artifact. Practically, since we do not know the corner-outlier-artifact candidate block, the four cases listed in Table 1 should be investigated. If two corner outlier artifacts appear at one cross-point, the artifacts are arranged on diagonally opposite sides because the corner outlier artifacts cannot be placed vertically or horizontally. In this case, the proposed method can detect both artifacts without additional computations, since we examine the four cases independently using the equations in Table 1.

To detect corner outlier artifacts based on the second observation, we examine whether the block satisfying the condition in Table 1 is flat or not. That is, when the candidate block is **A**, we examine flatness of block **A** with respect to the pixel including the corner outlier artifact using

$$A_{flat} = \sum_{i=2}^9 |a_1 - a_i| \quad (2)$$

where a_i is the pixel including the corner outlier artifact. We consider block **A** flat when A_{flat} is less than QP. For another block, we consider the block flat when the followings are less than QP, respectively.

$$B_{flat} = \sum_{i=2}^9 |b_1 - b_i|, \quad C_{flat} = \sum_{i=2}^9 |c_1 - c_i|, \quad D_{flat} = \sum_{i=2}^9 |d_1 - d_i| \quad (3)$$

where each capital and small letter denotes blocks and pixels, respectively. When the conditions of Table 1 are satisfied and (2) is less than QP, we regard the block as including a corner outlier artifact, and then the proposed filter is applied to the block in order to remove the artifact.

3 Proposed Filter to Remove Corner Outlier Artifacts

To remove the corner outlier artifacts, we propose a filter that updates pixels of the stair-shaped discontinuity using neighboring pixels. To reduce computational

complexity, we apply the proposed filter under the assumption that the edge has a diagonal direction instead of detecting the actual edge direction of neighboring blocks. This is reasonable because:

1. Corner outlier artifacts created by a diagonal edge are more noticeable than the artifacts created by a horizontal or vertical edge at a cross-point.
2. Various deblocking filters can remove corner outlier artifacts created by a horizontal or vertical edge.

According to the above assumption and the detection results obtained in Section 2, when block **A** includes a corner outlier artifact as shown in Fig. 2(b), the pixels in block **A** are replaced by

$$\begin{cases} a_1' = (b_1 + b_2 + c_1 + d_1 + d_2 + c_3 + d_3 + d_4) / 8 \\ a_2' = (a_2 + a_1' + b_1 + c_2 + c_1 + d_1 + c_4 + c_3 + d_3) / 9 \\ a_3' = (a_3 + b_3 + b_4 + a_1' + b_1 + b_2 + c_1 + d_1 + d_2) / 9 \\ a_4' = (a_4 + a_3' + b_3 + a_2' + a_1' + b_1 + c_2 + c_1 + d_1) / 9 \\ a_5' = (a_5 + b_5 + b_6 + a_3' + b_3 + b_4 + a_1' + b_1 + b_2) / 9 \\ a_9' = (a_9 + a_2' + a_1' + c_9 + c_2 + c_1 + c_8 + c_4 + c_3) / 9 \end{cases} \quad (4)$$

where each index follows that of Fig. 3. For blocks **B**, **C**, and **D**, the filtering methods are similar to (4) and actual equations are as follows: in the case where **B** includes a corner outlier artifact, as seen in Fig. 2(c), the pixels in block **B** are replaced by:

$$\begin{cases} b_1' = (a_1 + a_2 + d_1 + c_1 + c_2 + d_3 + c_3 + c_4) / 8 \\ b_2' = (b_2 + b_1' + a_1 + d_2 + d_1 + c_1 + d_4 + d_3 + c_3) / 9 \\ b_3' = (b_3 + a_3 + a_4 + b_1' + a_1 + a_2 + d_1 + c_1 + c_2) / 9 \\ b_4' = (b_4 + b_3' + a_3 + b_2' + b_1' + a_1 + d_2 + d_1 + c_1) / 9 \\ b_5' = (b_5 + a_5 + a_6 + b_3' + a_3 + a_4 + b_1' + a_1 + a_2) / 9 \\ b_9' = (b_9 + b_2' + b_1' + d_9 + d_2 + d_1 + d_8 + d_4 + d_3) / 9 \end{cases} \quad (5)$$

in the case where block **C** includes a corner outlier artifact, as seen in Fig. 2(d), the pixels in block **C** are replaced by:

$$\begin{cases} c_1' = (d_1 + d_2 + a_1 + b_1 + b_2 + a_3 + b_3 + b_4) / 8 \\ c_2' = (c_2 + c_1' + d_1 + a_2 + a_1 + b_1 + a_4 + a_3 + b_3) / 9 \\ c_3' = (c_3 + d_3 + d_4 + c_1' + d_1 + d_2 + a_1 + b_1 + b_2) / 9 \\ c_4' = (c_4 + c_3' + d_3 + c_2' + c_1' + d_1 + a_2 + a_1 + b_1) / 9 \\ c_5' = (c_5 + d_5 + d_6 + c_3' + d_3 + d_4 + c_1' + d_1 + d_2) / 9 \\ c_9' = (c_9 + c_2' + c_1' + a_9 + a_2 + a_1 + a_8 + a_4 + a_3) / 9 \end{cases} \quad (6)$$

and in the case where block **D** includes a corner outlier artifact, as seen in Fig. 2(a), the pixels in block **D** are replaced by:

$$\left\{ \begin{array}{l} d_1' = (c_1 + c_2 + b_1 + a_1 + a_2 + b_3 + a_3 + a_4) / 8 \\ d_2' = (d_2 + d_1' + c_1 + b_2 + b_1 + a_1 + b_4 + b_3 + a_3) / 9 \\ d_3' = (d_3 + c_3 + c_4 + d_1' + c_1 + c_2 + b_1 + a_1 + a_2) / 9 \\ d_4' = (d_4 + d_3' + c_3 + d_2' + d_1' + c_1 + b_2 + b_1 + a_1) / 9 \\ d_5' = (d_5 + c_5 + c_6 + d_3' + c_3 + c_4 + d_1' + c_1 + c_2) / 9 \\ d_9' = (d_9 + d_2' + d_1' + b_9 + b_2 + b_1 + b_8 + b_4 + b_3) / 9 \end{array} \right. \quad (7)$$

In each equation, the indices follow those of Fig. 3.

4 Simulation Results

The proposed method was applied to ITU test sequences at various bit-rates. To evaluate the proposed method, each test sequence was coded using the MPEG-4 verification model (VM) [6] with two coding modes: IPPP..., i.e., all frames of a sequence are inter-frame coded except the first frame, and I-only, i.e., all frames are intra-frame coded. In each mode, we applied the proposed filter to the reconstructed frames with none of the coding options being switched on and with the deblocking filter [2] being switched on, respectively. To arrive at a certain bit-rate, an appropriate quantization parameter was chosen and kept constant throughout the sequence. This can avoid possible side effects from typical rate control methods.

Table 2. PSNR results for IPPP... case

Sequence	QP	Bit-rate (kbps)	PSNR_Y (dB)			
			No filtering	No filtering + proposed filter	Deblocking	Deblocking + proposed filter
Hall Monitor	18	9.39	29.8728	29.8851	30.1957	30.2090
Mother & Daughter	16	9.45	32.2256	32.2469	32.3684	32.3894
Container Ship	17	9.8	29.5072	29.5082	29.7281	29.7291
Hall Monitor	9	24.29	34.0276	34.0325	34.1726	34.1835
Mother & Daughter	8	23.83	35.3303	35.3316	35.2686	35.2708
Container Ship	10	21.61	32.5701	32.5711	32.6003	32.6013
Foreman	14	46.44	30.6237	30.6252	31.0727	31.0739
Coastguard	14	44.68	29.0698	29.0763	29.1562	29.1586
Hall Monitor	12	47.82	33.8216	33.8236	34.0921	34.0948
News	19	47.21	31.2192	31.2202	31.3516	31.3528
Foreman	12	64.65	31.4786	31.4798	31.7587	31.7598
News	16	63.14	32.0648	32.0658	32.1233	32.1243

The simulation results for IPPP... coded sequences are summarized in Table 2. It can be seen that PSNR results for the luminance component are increased by up to 0.02dB throughout the sequences. Just a slight improvement in PSNR is obtained due to limitation of the area satisfying the filtering conditions. However, the proposed filter considerably improves subjective visual quality, particularly for sequences with apparent diagonal edges such as *Hall Monitor*, *Mother & Daughter*, *Foreman*, etc.

For emphasizing the effect of the proposed filter, partially enlarged frames of the first frame of *Hall Monitor* sequence under each method, i.e., the result of the proposed method without and with deblocking filter, respectively, are shown in Fig. 4.

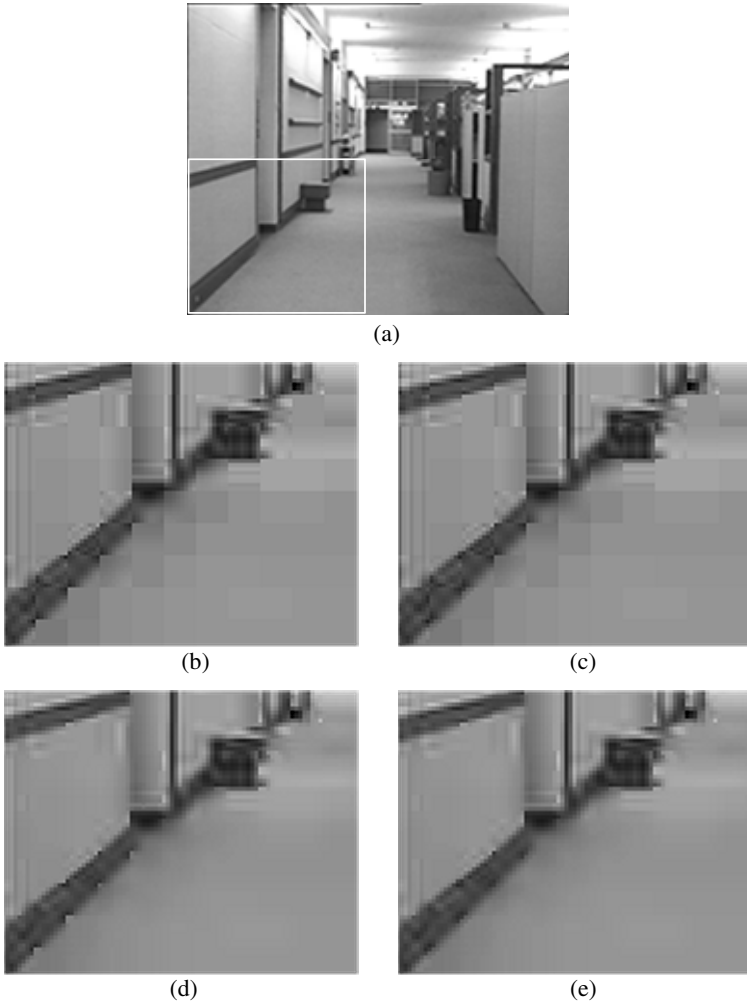


Fig. 4. Result images for *Hall Monitor* sequence (a) Original sequence (QCIF, QP=17), (b) and (d) Partially enlarged images of MPEG-4 reconstructed images without and with the MPEG-4 deblocking filter, respectively, (c) and (e) Partially enlarged images of the proposed method without and with the MPEG-4 deblocking filter, respectively

Table 3 shows the results for I-only coded sequences of *Hall Monitor* and *Foreman*. The results are similar to the IPPP... coded case. The proposed filtering conditions are satisfied at low QP for the sequences that include low spatial details

such as *Hall Monitor*, since each frame is relatively flat originally. On the other hand, the proposed filtering conditions are satisfied at relatively high QP for the sequences that include medium or high spatial details such as *Foreman*, since each frame is flattened at high QP. This tendency maintains in the sequences with or without the deblocking filter.

Table 3. PSNR results for I-only case

Sequence	QP	PSNR_Y (dB)			
		No filtering	No filtering + proposed filter	Deblocking	Deblocking + proposed filter
Hall Monitor	12	32.8170	32.8396	33.2223	33.2422
	17	30.5284	30.5381	30.9624	30.9848
	22	28.9326	28.9374	29.4025	29.4052
	27	27.6323	27.6344	28.1188	28.1230
Foreman	12	32.1830	32.1862	32.5741	32.5760
	17	30.0531	30.0606	30.5267	30.5316
	22	28.6406	28.6540	29.1635	29.1745
	27	27.5515	27.5684	28.1405	28.1526

5 Conclusions

The corner outlier artifacts are very annoying visually in highly compressed video although they appear in limited areas. To remove such artifacts, we have proposed a simple and effective post-processing method, which includes a detection method and a compensation filter. The proposed method, particularly in combination with the deblocking filter, further improves both objective and subjective visual quality.

Acknowledgments. This work was supported by the Korea Research Foundation Grant funded by the Korean Government (MOEHRD) (KRF-2004-005-D00164).

References

1. Park, H., Lee, Y.: A Postprocessing Method for Reducing Quantization Effects in Low Bit-rate Moving Picture Coding. *IEEE Trans. Circuits and Syst. Video Tech.*, Vol. 9. (1999) 161-171
2. Kim, S., Yi, J., Kim, H., Ra, J.: A Deblocking Filter with Two Separate Modes in Block-based Video Coding. *IEEE Trans. Circuits and Syst. Video Tech.*, Vol. 9. (1999) 156-160
3. Zhao, Y., Cheng, G., Yu, S.: Postprocessing Technique for Blocking Artifacts Reduction in DCT Domain. *IEE Electron. Lett.*, Vol. 40. (2004) 1175-1176
4. Wang, C., Zhang, W.-J., Fang, X.-Z.: Adaptive Reduction of Blocking Artifacts in DCT Domain for Highly Compressed Images. *IEEE Trans. Consum. Electron.*, Vol. 50. (2004) 647-654
5. Kaup, A.: Reduction of Ringing Noise in Transform Image Coding Using Simple Adaptive Filter. *IEE Electron. Lett.*, Vol. 34. (1998) 2110-2112
6. Text of MPEG-4 Video Verification Model ver.18.0. ISO/IEC Doc. N3908, (2001)

Content-Based Model Template Adaptation and Real-Time System for Behavior Interpretation in Sports Video

Jungong Han¹ and Peter H.N. de With^{1,2}

¹ University of Technology Eindhoven, P.O.Box 513, 5600MB Eindhoven
jg.han@tue.nl

² LogicaCMG, RTSE, PO Box 7089, 5605JB Eindhoven, The Netherlands

Abstract. In this paper, we present a *real-time* sports analysis system, which not only recognizes the semantic events, but also concludes the behavior, like player's tactics. To this end, we propose an advanced multiple-player tracking algorithm, which addresses two improvements on practical problems: (1) updating of the player template so that it remains a good model over time, and (2) adaptive scaling of the template size depending on the player motion. In this algorithm, we obtain the initial locations of players in the first frame. The tracking is performed by considering both the kinematic constraints of the player and the color distribution of appearance, thereby achieving promising results. We demonstrate the performance of the proposed system by evaluating it for double tennis matches where the player count and the resulting occlusions are challenging.

1 Introduction

Sports events gain wide interest and are among the most popular media attractions in the world today. Due to the the growth in hard-disk capacity, it is now possible to record about a thousand hours of video on one disk. Therefore, applications for organizing and analyzing the augmenting video data are emerging and may have a large market impact. Automatic sports analysis systems is one of those applications. However, for applications that should facilitate various users with different preferences, existing systems still cannot provide satisfactory results in most cases.

Significant research in the area of sports video analysis has been performed, which can be broadly divided into three stages. Earlier publications [1], have only focused on pixel and/or object-level analysis, which segments court lines and/or tracks the moving players. Evidently, these systems do not provide the semantic meaning of a sports game. The second generation of sports video analysis is based on the analysis and exploration of highlights of the game. In [2], the authors observe that the interesting events are often replayed in slow motion immediately after they occur. Such an algorithm can be applied to analyze various sports games, but it is impossible to provide sufficient understanding of

a sports game, since a viewer cannot deduce the complete story only in terms of the special event. Recently, researchers have paid more attention to event-based content-analysis systems [3]-[5], aiming at detecting predefined events that are considered most interesting in a particular sport genre. Object color and texture features are employed to detect events and parse a TV soccer program [3]. But for many applications, it is difficult to achieve satisfactory results by using such an algorithm, because the events have a complex semantic nature and there are no distinct relations between low-level features and the corresponding semantic concepts. Sudhir *et al.* [4] propose a tennis-video analysis system approaching a video retrieval application. It detects the court lines and tracks the moving-players, then extracts the events, such as base-line rally, based on the relative position between the player and the court lines. Unfortunately, its scene-level analysis model is rather limited, because only position information is employed to extract events. Kijak *et al.* [5] first define four types of view in tennis video, involving global, medium, close-up and audience, and then detect events like first-service failure in terms of the interleaving relations of these four views. This shot-based model does not take object behavior into account, so that it is not able to provide sufficient classification capabilities. In our previous work [9], we present a multi-level sports analysis system, which provides various services at three levels, such as pixel-level, object-level and scene-level. But this system fails to analyze a double-match tennis game, as its moving-player detection algorithm does not track multiple players simultaneously. To sum up, most existing systems, in particular tennis analysis, cannot analyze a sports game at multiple semantic layers, since they are unable to bridge the gap among the different layers. In addition, to the best of our knowledge, there is no system that can parse a double-match tennis game, as the player detection algorithms adopted by existing systems always fail at the multiple players case.

This paper attempts to solve the above problems, where we contribute on two aspects. Firstly, we present an automatic algorithm to track the *multiple* moving players, while addressing two specific problems. In our approach, we model the objects (players) by using a template technique such as in the work of [7]. The key to our solution is that we introduce a special content-based function to automatically adapt the template whenever it is needed, which better suits to the dynamics of sports. This overcomes the adaptation problems with conventional algorithms, such as periodic modification without linking to the content. Secondly, we provide a more advanced system that not only analyzes a tennis video at multiple layers, but also draws conclusions about player behavior, like the player's game tactics. This is profited from employing a three-dimensional (3D) camera model that bridges the gap between the video and the real-world in order to make more robust interpretations.

The paper is organized as follows, Section 2 introduces our advanced multiple players tracking technique. Section 3 describes the system for behavior interpretation. The experimental results on tennis video sequences are provided in Section 4. Finally, Section 5 presents conclusions.

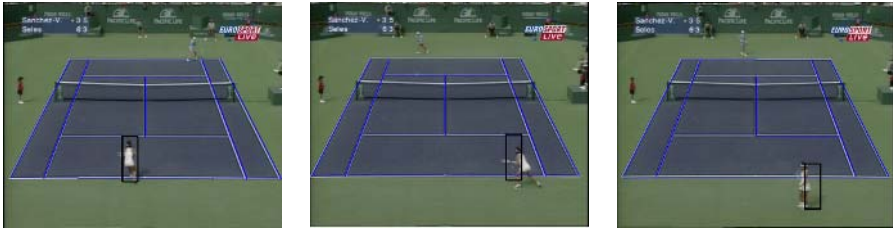


Fig. 1. The need of template update strategies. Left: initial location of a player in the first frame. Middle: update template every frame, where the template drifts away at the 32th frame. Right: the template is not updated and tracking eventually fails at the 81th frame.

2 Multiple Players Tracking System

To analyze the video at semantic level, it is required to track the players over a sequence. Substantial work has been done in this area. In [6], the tracking is performed by the maximization of a joint probability model taking both player’s motion and appearance into account. Needham *et al.* [1] has presented a framework for multi-player tracking, using a condensation-based approach. Each player being tracked is independently modeled, and the matching probability is calculated and used to select the best model fit for each player. To sum up, the previous algorithms have primarily focused on addressing the problems of regular human tracking, such as the technique of modeling the appearance of a target object. In other words, little attention is devoted to the specific problems caused by the nature of sports. Compared to human tracking in the surveillance application [7], tracking players over a sports video sequence, especially for obtaining accurate position of each player, is more difficult because of the following issues.

1. *Template update.* The players move rapidly, and show large silhouette deformations. This means that there are more drifting errors introduced into the template than those occurring with normal human tracking when updating the tracking template. In this case, the solution that either updates the template every frame (or every n frames) or no-update cannot prevent the template to steadily drift away from the target. See Fig. 1 for an example.
2. *Template-size scaling.* Another issue is that the players often move away or towards the camera with a high speed, thus the size of players in the image domain will change dramatically. For the applications like player behavior analysis, they require the player’s position with high accuracy. In other words, the template should be adaptively scaled with the size of the player body. Regularly, a naive solution adopted by existing systems is to change the template size with a fixed parameter.

2.1 Proposed Algorithms

Prior to introducing our new template-adaptation technique, we present a generic formal explanation of the tracking procedure. Assume there is a template for

the n^{th} frame, and the probability of the feature $\{u_i\}_{i=1\dots m}$ in the template is $\hat{T}_{n-1}(u_i)$, where u_i represents the color histogram distribution and i denotes the bin number of the histogram. Furthermore, we use the same method to model the probability of the target candidates, and represent them as $\hat{W}_n(u_i, P)$, where P denotes the center point of a candidate. The aim of template tracking is to find the best match to the template. Mathematically, finding the best match means maximizing the correspondence between the template and the candidates at arbitrary positions P , so that we compute

$$C_n = \arg \max_P \rho(\hat{W}_n(u_i, P), \hat{T}_{n-1}(u_i)). \quad (1)$$

Here, the term $\rho(\hat{W}_n(u_i, P), \hat{T}_{n-1}(u_i))$ is a metric to measure the matching degree between the template and the target candidates. In this paper, for the metric function ρ , we use a divergence-type similarity function because of its performance, namely the Bhattacharyya coefficient [7], which is defined as:

$$\rho(\hat{W}, \hat{T}) = \sum_i \sqrt{\hat{W}_n(u_i, P) \times \hat{T}_{n-1}(u_i)}. \quad (2)$$

(1) Similarity Function-Based Template Update Strategy

As already mentioned, due to the rapid changes of the objects, the template for the matching is drifting. Let us now consider this phenomenon using the previously specified metric function. Assume we extract a template $\hat{T}_1(u_i)$ in the first frame and we update it every frame, then the template tracking technique results in

$$\begin{aligned} C_2 &= \arg \max_P \rho(\hat{W}_2(u_i, P), \hat{T}_1(u_i)), \\ C_3 &= \arg \max_P \rho(\hat{W}_3(u_i, P), (\hat{T}_1(u_i) + e_2)), \\ C_n &= \arg \max_P \rho(\hat{W}_n(u_i, P), (\hat{T}_1(u_i) + e_2 \dots + e_{n-1})). \end{aligned} \quad (3)$$

Here, e_i represents the drifting error introduced into the template by the i^{th} frame ($e_1 = 0$). As seen in (3), each time the template is updated, drifting errors are introduced in the location of the template. With each update, these errors accumulate. If the accumulated drifting errors become too large and we still update the template, the target will be finally lost. In video surveillance applications, there are no large drifting errors introduced into the template because the human object moves with a normal speed and the body has no significant deformation. In this case, the template hardly drifts away from the target, even it is updated every frame. Based on the analysis above, we conclude that the template should better be updated when the drifting errors are small (and remain unchanged when having large errors). The experimental results of [8] (the

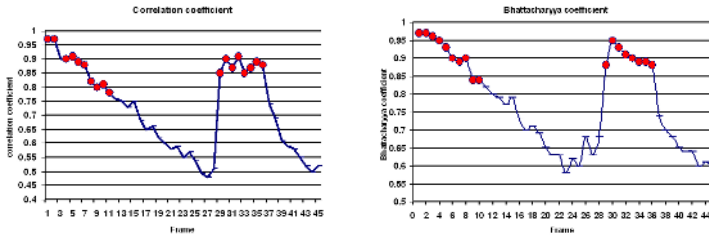


Fig. 2. An illustration of the correlation degree (Hor.: frame number, Vert.: correlation degree). The dots represent that the template is updated at that frame. Left: drifting errors among $\hat{T}_1(u_i)$ and target candidate. Right: Bhattacharyya coefficient of $\hat{T}_{n-1}(u_i)$ and target candidate.

left picture of Fig. 2) also support our conclusion. In the experiment, we have manually measured the drifting error using the correlation coefficient between the selected target candidate and the initial template $\hat{T}_1(u_i)$. Obviously, the drifting errors are inversely proportional to the correlation coefficient. In Fig. 2, the updates (dots) are always located at the frames having a higher correlation coefficient.

Now, the problem shifts to how to bridge the drifting error and the template update, since we cannot find an absolute rule between them from the statistical experiment in some cases. For example, a frame with a 0.5 correlation coefficient (relative bigger drifting errors) needs to update its template, while a frame that the coefficient is 0.6 (relative smaller drifting errors) may not require the template updating. A good template-update strategy should be capable of bridging the gap between the drifting errors and updating template. In [8], a smart approach is proposed that keeps the first template $\hat{T}_1(u_i)$ and uses it to correct the drift in $\hat{T}_{n-1}(u_i)$. To this end, a template matching procedure is employed using $\hat{T}_{n-1}(u_i)$ for the n^{th} frame. Based on it, a similar matching procedure is applied using $\hat{T}_1(u_i)$. The distance between the two positions obtained by the two matching procedures is adopted to determine template update. The computational cost of this algorithm is rather high, as it applies two tracking procedures to every frame. In this paper, we employ a *Similarity Function* (SF) to bridge the drifting errors and the template update, thereby eliminating the second template matching procedure in most frames [8]. The right picture of Fig. 2 portrays a SF curve that is obtained with (2). Note the main difference between the two curves in Fig. 2 is that the left one is based on the relations between the drifting errors and the template update, whereas the right one measures the similarity between the target candidate and $\hat{T}_{n-1}(u_i)$. From the results, we have found that the trends and characteristics of these two curves are similar. For this reason, the right curve can be applied to represent the drifting errors, and the following statistical results can help bridge the gap between the drifting error and the template update. (1) 98% frames update its template when the SF coefficient is bigger than 0.8, which means the errors is too small to drift the template. (2) If the SF coefficient is between 0.65 and 0.8, there is no a simple correlation

between the drifting error and the template update. (3) 99% frames never update the template in case of the SF coefficient is below 0.65.

Summarizing the above, we carry out our algorithm in two steps. Firstly, the use of (1) provides the best target candidate. Secondly, we selectively update the template. Suppose the Bhattacharyya coefficient of the current frame is B_n , and that of the previous frame is B_{n-1} . If $B_n > Th_1$, the template will be updated, and if $B_n < Th_2$, the template cannot be updated. For the frames where B_n is between Th_1 and Th_2 and if $|B_n - B_{n-1}| < 0.3$, the decision of updating the template in the current frame is the same as that of the previous frame. For other frames, we use the second template matching procedure [8] to decide.

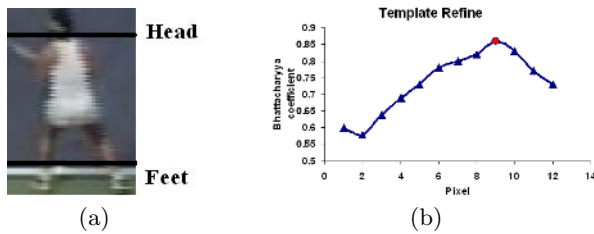


Fig. 3. Subregion-based template scaling. (a) Head and feet region, (b) Template refinement.

(2) Subregion-based Template Scaling Strategy

In this paper, two important subregions, head and feet (see Fig. 3(a)), are segmented and indicated by horizontal lines. As an aid, these subregions are used to compute the scale size of the template in the vertical direction. The main reason is that head and feet are relative rigid parts of the human body, which can be easily found. Once we have obtained the scale size in the vertical direction, the same scaling factor will be performed to the horizontal dimension. As an example, we now illustrate the detection of the head-region scaling factor when a player moves away from the camera.

More specifically, we define the histogram distribution probability of the head region extracted in the first frame as $\hat{H}_1(u_i)$. The initial head region segmented in the current frame is defined as $\hat{H}_n(u_i)$. We use Eq. (2) to measure the similarity between them. Subsequently, we obtain a new head region by shifting down the vertical boundaries of the initial head region with one line. Then the $\hat{H}_n(u_i)$ is updated, and the similarity is measured again. Repeating this shift procedure, we obtain a number of similarity coefficients, as shown in Fig. 3(b) (X represents over how many lines was shifted, and Y is the corresponding Bhattacharyya coefficient). The next step is to find the highest point (dot in Fig. 3(b)), which is the final location of the head region in the current frame. Compared to the conventional method, the adaptive scaling has made our proposal content-aware, and it enables an optimal fit to the body of the player.

2.2 Automatic Player Tracking System

In this section, we present an automatic tennis-player tracking system which is able to address the problems mentioned at the start of this paper, and which is based the following algorithmic steps.

1) **Player segmentation in the first frame.** The player segmentation algorithm [9] employs an object segmentation technique based on change detection, and summarizes several effective visual properties in the tennis video (e.g. uniform court color) to build a background, thereby achieving more accurate segmentation results.

2) **Template modeling.** A target template is represented by an rectangular region in our system. We use a histogram distribution to model the appearance of the template. Let $\{x_j^*\}_{j=1\dots n}$ be the *normalized* pixel locations in the region that define the template model. The region is normalized by the row and column dimensions h_x and h_y . An isotropic kernel with a convex and monotonic decreasing kernel profile $k(x)$ (see [7]) assigns smaller weights to pixels further away from the center. The use of such weights increases the robustness of the density estimation, since the peripheral pixels are often affected by occlusions. The function $b : R^2 \rightarrow \{1\dots m\}$ maps the pixel at location x_j^* to the index $b(x_j^*)$ of its bin in the quantized histogram space. The probability of the feature $\{u_i\}_{i=1\dots m}$ in the template model is then computed as

$$\hat{T}_1(u_i) = c \sum_j k(\|x_j^*\|^2) \delta[b(x_j^*) - u_i], \quad (4)$$

where δ is the Kronecker delta function. The normalization constant c is derived by imposing the condition that $\sum_i \hat{T}_1(u_i) = 1$. The same method can be used to model the head and feet region.

3) **Target candidates modeling.** Let $\{x_j\}_{j=1\dots n}$ be the normalized pixel location of the target candidate, centered at P in the current frame. Using the same method, we can model the target candidates.

4) **Moving-players tracking.** Equations (1) and (2) can track the template over a sequence. During the tracking procedure, our template-update strategy and refinement strategy are employed. Once the template is updated, the template will be remodeled using Eq. (4). When occlusions occur, the model combining the player's kinematic constraints and the color distribution of appearance [6] ensures our system tracks moving players well.

3 System for Real-Time Behavior Interpretation

Fig. 4 shows the architecture of the proposed system, which consists of four modules, each being briefly explained below.

1. *Playing-frame detection.* A tennis sequence not only includes scenes in which the actual play takes place, but also breaks or advertisements. Since only the playing frames are important for the subsequent processing, we efficiently extract the frames showing court scenes for further analysis [9].

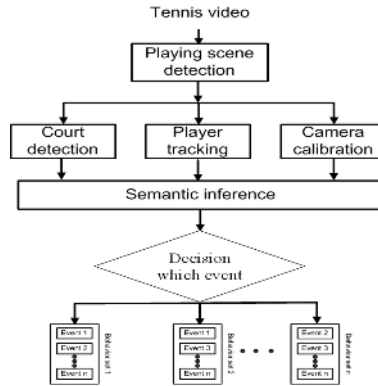


Fig. 4. Architecture of the real-time analysis system

2. *Court detection and camera calibration.* Court information, including size, shape and location, is an important aid to analyze the tennis game. To deduce the semantic meaning from the position and movements of the players, their position has to be known in real-world coordinates. However, pixel-level image-processing algorithms will only give the player positions in image coordinates, which are physically meaningless. To transform these image coordinates to physical positions, a camera-calibration algorithm has to be applied [10].
3. *Moving player segmentation and tracking.* We use our technique introduced before to position the players in the image domain, and then map it to the 3-D domain.
4. *Scene-level event classification and behavior interpretation.* This unit first selects several visual cues that enable to describe important events at high level. Then it uses an weighted linear combination model [9] and also game-specific contextual information to recognize the events of the tennis game. Different from [9], a behavior-interpretation module is added to our system. *Behavior* is defined as a sequence of events, with or without temporal constraints on the occurrence order of the events. Behavior analysis can be as simple as the detection of a single event, e.g., a player is approaching the net, or it can be a complex sequence of multiple events, e.g. both two players of a team are speeding up towards the net after one of them gave a service. Given the context of the tennis game, the player behavior (tactics) may be analyzed by defining behavior based on a sequence of events. For example, a standard “both up” tactics can involve two procedures (events), where the first is a successful service and the second is an event where two players from one team are immediately approaching the net.

4 Experimental Results

To evaluate the performance of the proposed algorithm, we tested our system on four tennis videos recorded from three different tennis matches (in total more

than 30 minutes). The new tracking system was applied to these video sequences, achieving an accurate 98% correct tracking rate. Sample frames are shown in Fig. 5, which demonstrates the tracking capability of our technique in case of occlusion. We compared our template-update algorithm with the method of [8] using 240 frames (720×576 TV resolution). These two methods are capable of finding the target over the sequence, but the computational cost is different. The total time consumed by our algorithm on a P-IV 3GHz PC is 52.17s, while the time required for the algorithm of [8] under the same conditions is 61.54s. This means that our system saves 15% computation time, while at the same time obtaining a higher quality. In addition, we tested our content-aware template scaling technique and compared it with the conventional technique that changes the template size with a fixed parameter, both under the same conditions. From Fig. 6, we can find our template-scaling approach is better than the conventional one. With our system the player is fully enclosed, while the conventional system takes an inappropriate template size.

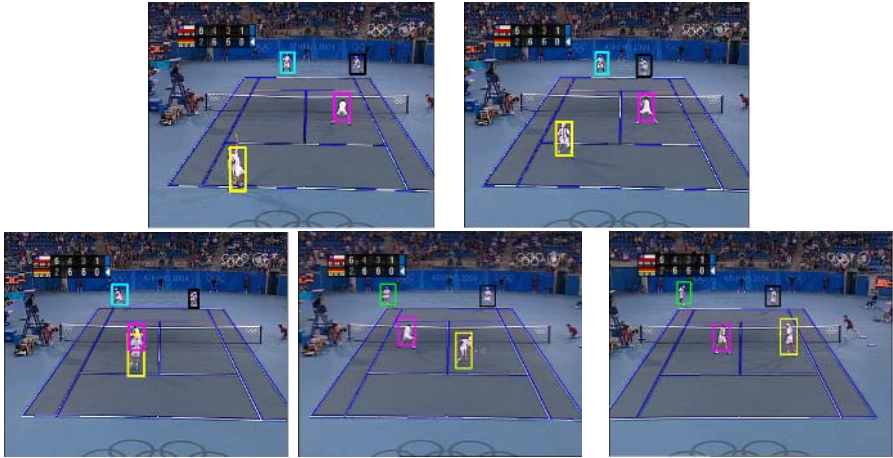


Fig. 5. Player tracking results for double-match game

We embedded the multiple-player tracking algorithm into our tennis analysis system, and obtained a 90% rate on event extraction. Fig. 7 shows the user interface of our tennis video analysis software. From this user interface, the viewer can find analysis results at three different levels. At the pixel level, several key objects are segmented and indicated. At the object level, the moving objects are tracked in the 3-D domain (at the right side). At the scene level (as opposed to [9]), the system not only detects the important events, but also indicates the behavior of each team, e.g., one team has “both up” tactics, while the other team is defending by “both back” tactics at the baseline. Table 1 shows the performance evaluation results of our system. It can be seen that our system achieves a real-time or near real-time performance (with a P-IV 3GHz PC).

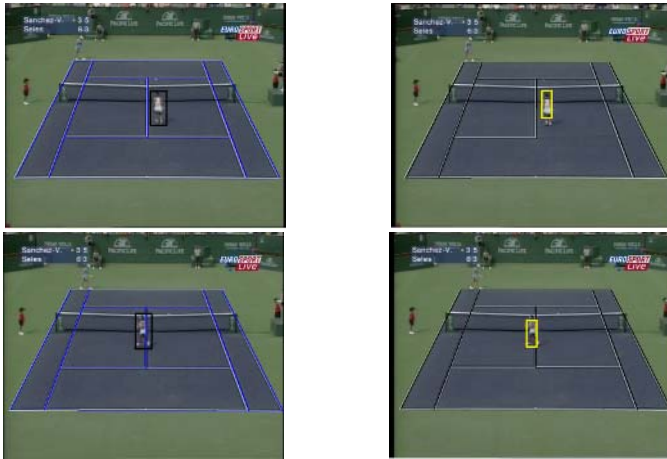


Fig. 6. Adaptive template scaling. Left: our method. Right: scaling with fixed parameter, leading to inappropriate sizes.

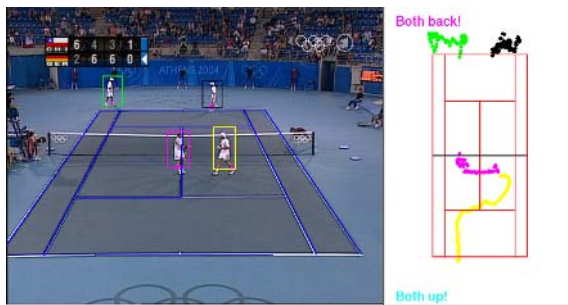


Fig. 7. Results of our analysis system. The left image shows the detected court and players. At the right is the real-word court model, where the trajectory of each player is visualized, as well as the player's behavior.

Table 1. System evaluation

Type	Resolution	Event detection	System efficiency
Clip1 single match	320×240	87%	6 frames/s
Clip2 single match	720×576	91%	3.5 frames/s
Clip3 double match	720×576	88%	3 frames/s

5 Conclusions

We have presented an automatic multi-player sports analysis system, which operates at various levels and can visualize features up to the player tactics. The

system has high algorithmic efficiency and processes in near real-time. A main improvement as compared to [9], is that we can now analyze the double-game of the tennis match. Here, two major contributions are realized. First, we designed a multiple-player tracking system, which can automatically update the template whenever it is needed and also scale the template-size according to dynamics of the player body. Second, we conclude the player behavior, e.g., player's tactics, based on recognizing high-level events and the game contextual information. The detection performance of the system for the double-game condition is in the order of 90% at a 3-6 frame/s rate. Although having obtained good results, experiments have revealed that the template-refinement algorithm is not sufficiently robust in case of noisy subregions and needs further optimization.

References

1. C. Needham and R. Boyle, "Tracking multiple sports players through occlusion, congestion and scale", Proc. 12th British Machine Vision Conference (BMVA) 2001, Vol. 1, pp. 93-102, 2001.
2. H. Pan, P. Beek and M. Sezan, "Detection of slow-motion replay segments in sports video for highlights generation," Proc. ICASSP2001, Salt Lake City, UT, May 2001.
3. Y. Gong, L. Sin, C. Chuan, H. Zhang, "Automatic parsing of soccer programs", Proc. IEEE Int. Conf. Mult. Comput. Syst., pp.167-174, 1995.
4. G. Sudhir, C. Lee and K. Jain, "Automatic classification of tennis video for high-level content-based retrieval", Proc. IEEE international workshop on content based access of image and video databases, pp. 81-90, 1998.
5. E. Kijak, L. Oisel and P. Gros, "Temporal structure analysis of broadcast tennis video using hidden Markov models ", Proc. SPIE Storage and Retrieval for Media Databases 2003, pp.289-299, January 2003.
6. J. Kang, I. Cohen and G. Medioni, "Soccer player tracking across uncalibrated camera streams", Proc. IEEE Int. Worksh. Visual surveillance and perform. eval. of tracking and surveillance, October 2003.
7. D. Comaniciu, V. Ramesh and P. Meer, "Kernel-based object tracking", IEEE Trans. Pattern Analysis Machine Intell., Vol. 25, No. 5, pp. 564-575, 2003.
8. I. Matthews, T. Ishikawa and S. Baker, "The template update problem", IEEE Trans. Pattern Analysis Machine Intell., Vol. 26, No. 6, pp. 810-815, 2004.
9. J. Han, D. Farin and P.H.N. de With, "Multi-level analysis of sports video sequences", Proc. SPIE Multimedia content analysis management and retrieval, San Jose (CA), Vol. 6073, No. 607303, pp. 1-12, January 2006.
10. D. Farin, J. Han and P.H.N. de With, "Fast camera-calibration for the analysis of sports sequences", Proc. IEEE Int. conf. Mult. Expo (ICME 2005), Amsterdam, pp. 482-485, July 2005.

New Approach to Wireless Video Compression with Low Complexity

Gangyi Jiang^{1,2}, Zhipeng Jin^{1,2}, Mei Yu^{1,2}, and Tae-Young Choi³

¹ Faculty of Information Science and Engineering, Ningbo University,
Ningbo 315211, China

² National Key Laboratory of Machine Perception, Peking University, 100871, China

³ Division of Electronics Engineering, Ajou University,
Suwon 442-749, Korea

Abstract. Because of limitation of low power and computational ability in mobile video devices, it is significant to develop energy-efficient wireless video compression methods for mobile video systems. In this paper, a new approach to low complexity wireless video compression is proposed, based on Wyner-Ziv coding theorem. The proposed method encodes video only by intracoding techniques and detection of regions of interest, without the complicated motion estimation and compensation in the mobile video terminal. Thus, the computational burden is obviously reduced and the requirement of low power can be satisfied in mobile video devices. Experimental results show that the proposed method is quite effective.

1 Introduction

The rapid growth in wireless video applications has resulted in spectacular strides in the progress of wireless communication systems. However, the stringent energy constraints of mobile devices and the high error rates of wireless channels still pose significant barriers in the deployment of wireless video applications. Current video coding standards, such as MPEG-x or H.26x, mainly rely on the powerful hybrid block-based motion compensation and discrete cosine transformation (MC/DCT) architecture, which account for a major share of the coding gain in rate-distortion (RD) performance. Because of the heavy computing burden of the motion estimation and compensation task in these video compression standards, the encoder is 5 to 10 times more complex than the decoder^[1-3]. This asymmetry in complexity is well-suited for broadcasting or for streaming video-on-demand systems where video is compressed once and decoded many times. But for wireless mobile terminals, such as wireless video sensors for surveillance, mobile camera phones, and networked camcorders, which have the limits of low power, storage and computational ability, the traditional video coding system is not suitable, since low complexity in video coding and transmission is required.

Wyner-Ziv Theorem on source coding with side information available only at the decoder suggests that an asymmetric video codec, where individual frames are encoded separately, but decoded conditionally (given temporally adjacent frames) could

achieve similar efficiency^[2]. Fig.1 shows a scheme for wireless mobile video communication, where the mobile video devices only perform low complexity encoding and decoding. From pixel-domain to DCT domain, Girod et al. accomplished a great deal of research on Wyner-Ziv video coding^[4-8]. Based on their scheme^[8], an improved scheme of wireless video coding with low complexity is proposed in this paper. A simple and efficient method of region of interest (ROI) segmentation is given through comparing the quantized DCT coefficients with those of previous key frame, so that coding of non-ROI, such as the background and smooth regions, can be avoided in the encoder. Experimental results show that the proposed scheme outperforms Girod's hash-based Wyner-Ziv video coding scheme in RD performance and computational burden.

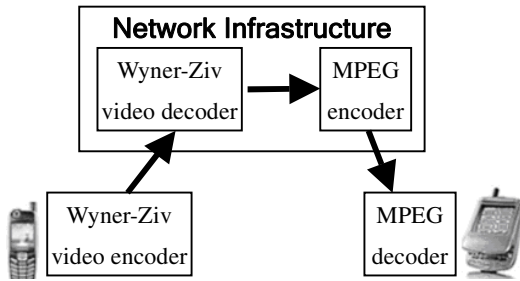


Fig. 1. Framework for wireless mobile video communication

2 New Approach to Wireless Video Compression with Low Complexity

Aaron and Girod presented a hash-based Wyner-Ziv video coding scheme in [6], where a blockwise DCT is applied to the Wyner-Ziv frame W . Then transform coefficients are divided into low frequency and high frequency sets. The low frequency coefficients are compressed with rate compatible punctured turbo code (RCPT)^[9]. The RCPT combined with a feedback loop, and the parity bits produced by the turbo encoder are stored in a buffer, which transmits a subset of these parity bits to the decoder upon the request. The high frequency components of the frame are treated as the hash code, and if sent, are compressed by efficient run-length coding and are used at the decoder in the inverse transform and in estimation the motion.

The decoder generates the side information frame \hat{W} , as an estimate of W , from the previous frame and the high frequency bits. For a given block of the current frame, if no high frequency bits are sent, the co-located block from the previous frame is used as the side information. If the high frequency bits are sent, the decoder reconstructs these coefficients and utilizes them in a motion search to generate the best side information block from the previous frame.

The turbo decoder uses the received subset of parity bits and the side information to decode the current bits. If the decoder cannot reliably decode the bits, it requests

additional parity bits from the encoder buffer through feedback. The requesting and decoding process is repeated until an acceptable probability of bit error is guaranteed.

In Grid's scheme, additional parity bits transmission controlled by the feedback loop is the key to gain good RD performance. But the feedback will bring heavy computing burden and delay, and it does not agree with the original intention of the low-complexity video encoding yet. In our experiments, the quantized DCT coefficients of the background and smooth regions in adjacent frames are discovered to be very approximate or even the same. By comparing the quantized DCT coefficients of current block in current Wyner-Ziv frame with that of the block at the same position in the previous key frame, the bit rate for encoding the background and smooth regions can be saved. So the regions, which disagree with the previous key frame, are referred to as the 'region of interest (ROI)', and the process of detecting these regions as 'ROI extraction', just as follows.

A quantized DCT coefficient difference (QDCD) operator is used to extract these regions of interest. QDCD is simple and accurate in segmenting the actual ROIs even though it might fail in few cases. More accurate ROI extraction scheme can increase the RD performance at the cost of increase in encoding complexity.

First, the difference d_i of each quantized DCT coefficient is defined as

$$d_i = |q_i - h_i|, \quad (1)$$

where q_i is the current quantized DCT coefficient of the current Wyner-Ziv frame, h_i is the quantized DCT coefficient of the previous key frame stored in encoder buffer as hash code.

Then, 'ROI extraction' is carried out according to difference d_i

$$\begin{aligned} & \text{if } \max(d_i) > 1, \text{ ROI block} \\ & \text{else if } \sum_{i \in N} R_i d_i \geq T, \text{ ROI block} \quad , \\ & \text{else background or smooth block} \end{aligned} \quad (2)$$

where R_i is the weight, the value of which is different according to the Zig-Zag scan positions. The weights of low frequency coefficients are larger than that of the high frequency coefficients. $i \in N$, and N represents the number of quantized DCT coefficients that hash code contains.

The ROI extraction scheme is simple but efficient. Compared with hash-based Wyner-Ziv coding, it doesn't increase additional computational complexity. Moreover, in comparison with storing the original pixel values, storing the quantized DCT coefficients requires less memories and computations. The ROI blocks in current Wyner-Ziv frame extracted by the proposed scheme are shown in Fig.2, in which it is obvious that lots of background and smooth regions are skipped from encoding.

The proposed ROI-based Wyner-Ziv video coding scheme is shown in Fig.3. As shown in the figure, the first three AC coefficients in Zig-zag scan in Fig.4 are considered as Part-1 (i.e., low coefficients), and encoded with PCCC (parallel

concatenated convolutional code); while the rest are regarded as Part-2, and encoded through CAVLC coding. It is noted that the Part-2 plays the same role as the hash code in the Grid's scheme, which is used at the decoder to estimate the motion information and obtain side information with high quality.

For each block of current Wyner-Ziv frame, the distance from the corresponding hash code of the previous key frame is calculated. The distance D is the weighted SAD between the subsampled version of the current block and the collocated quantized samples of the previous key frame, and defined by



Fig. 2. The ROI blocks in the current Wyner-Ziv frame extracted by the proposed scheme (the 11th frame of salesman)

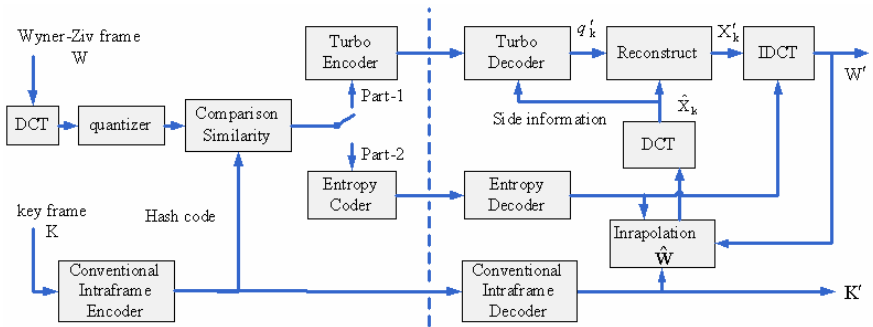


Fig. 3. ROI-based DCT domain Wyner-Ziv video coding

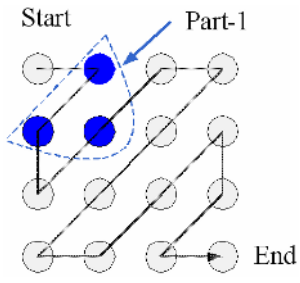


Fig. 4. Zig-Zag scan and division of DCT coefficients in 4x4 block

$$D = \sum_{i \in N} R_i d_i = \sum_{i \in N} R_i |q_i - h_i|. \quad (3)$$

1) If the distance D is smaller than a threshold T_1 , the block is considered as a non-ROI region, and the block is copied from the co-located block in the previous decoded key frame.

2) If the distance D of coefficients of Part-1 exceeds a threshold T_2 , these coefficients are encoded by PCCC, and the parity bits after 1/2 punctured are sent to the decoder directly, without stored in the encoder buffer. The feedback loop is not used in the proposed scheme in order to reduce the encoding complexity further.

3) If the distance D of coefficients of Part-2 exceeds a threshold T_3 , these coefficients are encoded with the CAVLC. If there is no channel error in video bitstream transmission, the decoder at the receiver can reconstructs these coefficients independently. In the scheme, the decoded quantized DCT coefficients of Part-2 are important to motion estimation at the decoder.

4) If the coefficients in Part-1 have been coded, the decoder uses the reconstructed Part-2 coefficients as a characterization of the original blocks to perform motion estimation in the previous decoded key frame, and choose the best matching block as the side information.

3 Experimental Results

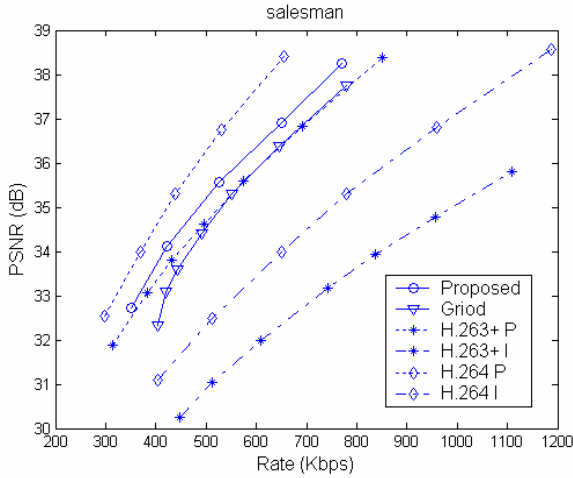
The proposed scheme is tested with salesman and container QCIF video sequences (at 30fps and with a total of 100 frames). For encoding Wyner-Ziv frame, 4x4 DCT transform is used and each coefficient is quantized with a uniform scalar quantizer. The turbo encoder is composed of two identical RSC encoders, and the corresponding

generator matrix is $\begin{bmatrix} 1 & \frac{1+D+D^3+D^4}{1+D^3+D^4} \end{bmatrix}$. The parity bits after 1/2 punctured is sent

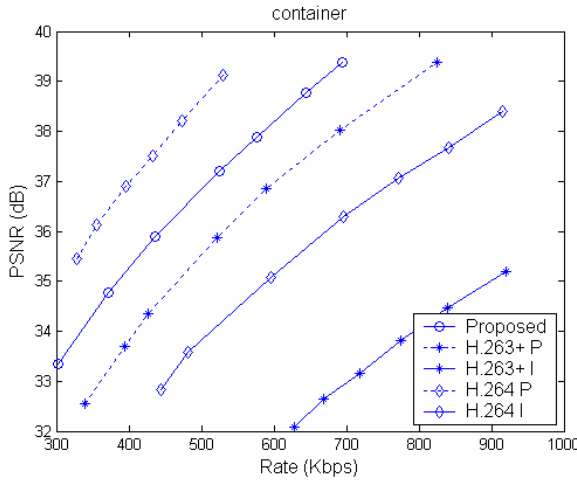
to the decoder directly, not stored in the encoder buffer.

Fig.5 shows the RD performances of the proposed scheme with respect to salesman and container test sequences. The figures show the total bit rate and average PSNR of both the key frames and the Wyner-Ziv frames. RD performance of the proposed scheme is compared with that of H.263+ video coding, H.264 video coding, and Grid's scheme^[6]. For justice, when H.263+ and H.264 interframe coding are carried out, the length of GOP is set to 2 (that is I-P-I-P, the same as the paper [6]). Besides, the SearchRange is chosen to be 16, and FME is used. The quantization parameters for I frames is the same as for P frames. The curve of 'Grid' comes from his paper [6], and the curve of 'Proposed' denotes the proposed scheme.

From Fig.5(a), it is seen that, for salesman sequence, the proposed scheme outperforms Grid's scheme and H.263+ interframe coding about 0.5dB at high bit rate. At lower bit rate the proposed scheme attains even 1.0dB more than Grid's scheme. Fig.5(b) gives similar results. For container sequence, the proposed scheme outperforms H.263+ interframe coding about 1.5dB on average.



(a) RD performance of salesman



(b) RD performance of container

Fig. 5. Rate-distortion (RD) performance

It should be noted that, although the H.264 interframe coding achieves the best RD performance just as shown in Fig.5, it is at the cost of huge computation complexity.

In order to achieve encoding with low-complexity, Wyner-Ziv coding adopts an intraframe encoder combined with an interframe decoder structure; therefore the encoding computation burden is much lower than any other interframe encoding systems. In addition, by taking advantage of effective ROI extraction, the background and smooth regions are avoided to be coded in the proposed scheme, so that the bit rate and encoding time are saved. Besides, the feedback loop is removed in the proposed scheme, which reduces the computational complexity further.

The total encoding time (for key frame and Wyner-Ziv frame) of the proposed scheme is given in Fig.6, in comparison of the total time of H.264 interframe coding. The experimental parameters are the same with the previous, the operating system of the PC is Windows2000, CPU is Pentium IV 3G, with 512M RAM, and the compiling software is Visual C++6.0. The experimental results indicate that, for salesman and container, the total encoding time of the proposed scheme is only 29.05% and 28.37% of that of H.264 interframe coding, respectively. So, no matter in the RD performance or encoding computational complexity, the proposed scheme is superior to Grid's hash-based Wyner-Ziv video coding.

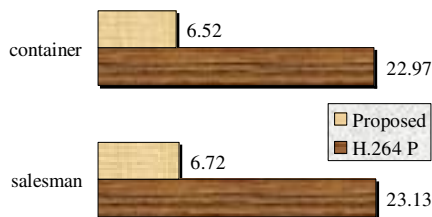


Fig. 6. Total encoding time of 100 frames (sec)



Fig. 7. Decoded Wyner-Ziv frames

Fig.7 shows the decoded images of Wyner-Ziv frame with respect to the proposed scheme. For the 95th frame of salesman (a Wyner-Ziv frame), it takes 4344bits and PSNR of the reconstructed image is 36.96 dB. For the 95th frame of container (also Wyner-Ziv frame), it takes 6512bit, and the corresponding PSNR is 36.11dB.

4 Conclusions

Based on the Grid's scheme, a ROI-based Wyner-Ziv video coding scheme with low encoding complexity is proposed. To save the bit rate, the proposed scheme detects ROI region by using the quantized DCT coefficients, so that the background and smooth regions need not to be coded. In addition, there is no feedback loop in the

proposed scheme, therefore the computational requirement is reduced further. Experimental results show that the proposed scheme outperforms Girod's hash-based Wyner-Ziv coding.

In future work, low-density parity-check (LDPC) code is used instead of turbo codes, and the probability of bit error at decoder will be reduced. In addition, a reconstruction filter is also expected to be studied, so as to reduce the blocking distortion of DCT transform and improve the subjective quality of coded images.

Acknowledgment

This work was supported by the Natural Science Foundation of China (grant 60472100), the Natural Science Foundation of Zhejiang Province (grant RC01057, 601017, Y105577), and the Ningbo Science and Technology Project of China (grant 2003A61001, 2004A610001, 2004A630002), and the Zhejiang Science and Technology Project of China (Grant 2004C31105).

References

1. Ostermann, J., Bormans, J., List, P., et al.: Video Coding with H.264/AVC: Tools, Performance, and Complexity, *IEEE Circuits and Systems*, vol.4, no.1, (2004) 7-28.
2. Jiang, G., Shao F., Yu, M., et al.: Efficient Block Matching for Ray-Space Predictive Coding in Free-Viewpoint Television Systems, *Lecture Notes in Computer Science, LNCS 3980*, (2006) 307-316.
3. Yu, M., Jiang, G., Li, S., et al.: New Approach to Complexity Reduction of Intra Prediction in Advanced Multimedia Compression, *Lecture Notes in Computer Science, LNCS 3980*, (2006) 317-325.
4. Girod, B., Aaron, A., Rane, S., Rebollo-Monedero, D.: Distributed video coding, *Proceedings of the IEEE*, vol.93, no.1, (2005) 71-83.
5. Aaron, A., Girod, B.: Wyner-Ziv video coding with low-encoder complexity, *Proc. Picture Coding Symposium, PCS 2004*, San Francisco, CA, (2004).
6. Aaron, A., Zhang, R., Girod, B.: Wyner-Ziv coding of motion video, *Proc. Asilomar Conference on Signals and Systems*, Pacific Grove, CA, (2002).
7. Aaron, A., Rane, S., Setton, E., Girod, B.: Transform-domain Wyner-Ziv codec for video, *Proc. Visual Communications and Image Processing, VCIP-2004*, San Jose, CA, Jan. 2004.
8. Aaron, A., Rane, S., Girod, B.: Wyner-Ziv video coding with hash-based motion compensation at the receiver, *Proc. IEEE ICIP-2004*, Singapore, (2004).
9. Rowitch, D., Milstein, L.: On the performance of hybrid FEC/ARQ systems using rate compatible punctured turbo codes, *IEEE Trans. on Commun.*, vol.48, no.6, (2000) 948-959.

Fast Multi-view Disparity Estimation for Multi-view Video Systems

Gangyi Jiang^{1,2}, Mei Yu^{1,3}, Feng Shao^{1,3}, You Yang^{1,2}, and Haitao Dong¹

¹ Faculty of Information Science and Eng., Ningbo University,
Ningbo, 315211, China

² Institute of Computing Technology, Chinese Academy of Science,
Beijing, 100080, China

³ National Key Laboratory of Machine Perception, Peking University,
Beijing, 100871, China

Abstract. Disparity estimation can be used to eliminate redundancy among different views in multi-view video compression to obtain high compression efficiency. However, the problem of high computational complexity in disparity estimation, which limits real-time applications of multi-view systems, needs to be solved. In this paper, a novel fast multi-view disparity estimation algorithm based on Hadamard similarity coefficient for multi-view video coding is proposed by using prediction of initial search point, selection of reference view, determination of the best disparity vector, and strategies of search stop. Experimental results show that the proposed algorithm can significantly reduce the computational complexity in multi-view disparity estimation.

1 Introduction

Convergence of technologies from multimedia, telecommunications, computer graphics, and related fields enables the development of applications that significantly extend the sensation of classical 2D video. The new types of applications allow the user to freely choose a viewpoint of a visual scene or/and provide a 3D depth impression of a visual scene^[1-3], such as free viewpoint video systems and 3DTV. Multi-view video coding (MVC) plays an important role to free viewpoint video systems, which can provide video information with respect to different angle of a scene. MPEG has already made attention to MVC in 3D audio-visual systems^[1,4].

There is evident redundancy between views, while disparity estimation is a key technology to eliminate the redundancy between views in MVC^[4-6]. Just like motion estimation to remove temporal redundancy in mono-view video coding, disparity estimation can be used to eliminate effectively the redundancy among views in MVC. However, disparity estimation is heavy computational burthen in whole system. Kimata has proposed multi-reference views based disparity estimation algorithm^[6], which is able to enhance the rate-distortion performance, but its computational complexity is doubled due to the multi-reference frame searching. Lopez proposed block-based illumination compensation and search techniques^[7], but its computational complexity is still a problem. In this paper, a novel fast disparity estimation algorithm

based on Hadamard similarity coefficient is proposed, in which some techniques, including prediction of initial search point, selection of reference view, determination of the best disparity vector and strategies of search stop, are integrated so as to reduce the computational complexity of disparity estimation. Experimental results show that the proposed algorithm can significantly reduce the computational complexity in multi-view disparity estimation.

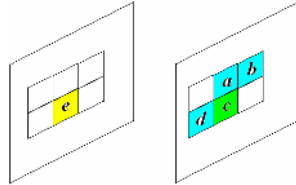


Fig. 1. Definition of two kinds of neighboring blocks

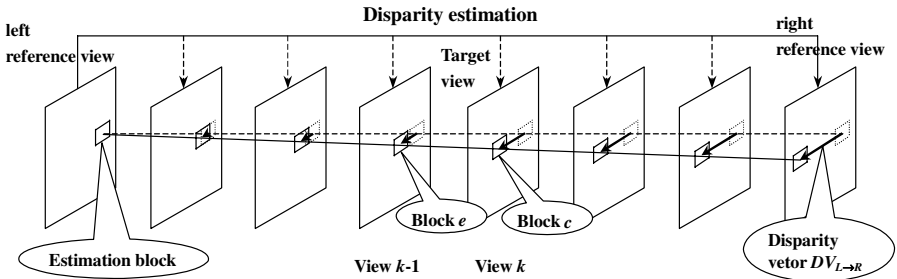


Fig. 2. Disparity interpolation of multi-view images

2 Characteristic Analysis of Neighboring Blocks

Multi-reference disparity estimation can improve efficiency of MVC, but its computational complexity will be doubled when the number of its reference views increases. To reduce the computational complexity in disparity estimation, fast algorithms are needed. In this paper, definitions of two kinds neighboring blocks are given in Fig.1, where blocks *a*, *b*, and *d* are the encoded neighboring blocks of the current block *c* in the same frame, called as intra-view neighboring blocks. Block *e*, on the other hand, is the predicted block of the current block *c* in the adjacent view, called as inter-view neighboring block. Fig.2 shows disparity interpolation of multi-view images, in which a relationship between the block *c* and block *e* may be determined by disparity interpolation.

In this section, Hadamard similarity coefficient is defined to describe the relationship between neighboring blocks. Hadamard transform is an orthogonal transform with very low computational complexity, since only addition and subtraction are needed, and the characteristics of its transform coefficients are similar to those of DCT. Fig.3 gives statistical results of Hadamard coefficients of blocks

with the size of 8×8 . It is seen that for smooth regions of image the energy concentrates at the position of $(0,0)$, $(0,2)$, $(2,0)$, while for close-grained regions of image the energy mainly concentrates at the position of $(0,0)$, $(0,2)$, $(2,0)$, $(0,4)$, $(4,0)$ and $(4,4)$.

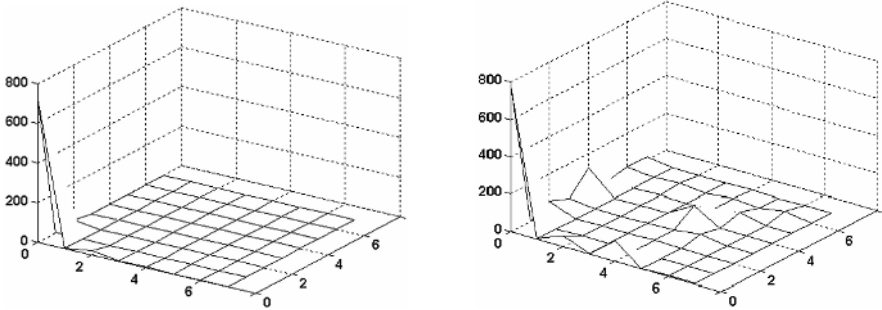


Fig. 3. (a) Statistical result of smooth region, (b) Statistical result of close-grained region

According to the above statistical results, similarity coefficient of Hadamard coefficients is defined. Let $H(i,j)$ be Hadamard coefficient at position (i,j) , $C_{i,j} = |H_1(i,j) - H_2(i,j)|$ be the absolute difference between Hadamard coefficients of two blocks, $S_{i,j} = |H_1(i,j) + H_2(i,j)|$ be the absolute sum of Hadamard coefficients of two blocks. For intra-view neighboring block, the Hadamard similarity coefficient is defined by formula (1), in this case the smooth region is usually considered.

$$R = (C_{0,0} + C_{0,2} + C_{2,0}) / S_{0,0} \quad (1)$$

When inter-view neighboring block is considered, the Hadamard similarity coefficient is defined as the following to describe characteristic of close-grained region.

$$R = (C_{0,0} + C_{0,2} + C_{2,0} + C_{0,4} + C_{4,0} + C_{4,4}) / (S_{0,0} + S_{4,4}) \quad (2)$$

It is noted that it is not necessary to calculate all Hadamard coefficients, but only the coefficients at the positions of $(0,0)$, $(0,2)$, $(2,0)$, $(0,4)$, $(4,0)$ and $(4,4)$, so that the computational burden is very slight.

3 A New Fast Multi-view Disparity Estimation Algorithm

Based on characteristic analysis of neighboring blocks, a novel fast multi-view disparity estimation algorithm is proposed. First, disparity maps between the leftmost and rightmost views are estimated, for which the leftmost and rightmost views are used as reference view mutually. The obtained disparity map $\{DV_{L \rightarrow R}\}$ and $\{DV_{R \rightarrow L}\}$ are as the initial ones. For disparity estimation of the current block c in the current view k , intra-view Hadamard similarity coefficients $\{R_{na}, R_{nb}, R_{nd}\}$ of the blocks a , b and d are first calculated, let R_n be the minimum of them, then R_n is compared with a similarity threshold R_T . If $R_n < R_T$, the two blocks are considered to

be similar, called as intra-view similar block, so the disparity vector of neighboring block with respect to R_n is used as estimated disparity vector $DV_c(0)$ of the current block c , moreover, the two blocks have the same reference view. Let SAD_a , SAD_b and SAD_d be the sum-of-absolute-difference (SAD) of the three neighboring blocks, and $SAD_c(0)$ of the current block c be the SAD value obtained by using $DV_c(0)$ and the corresponding reference view. Let $med\{\}$ denote a median operation. If $SAD_c(0) \leq med\{SAD_a, SAD_b, SAD_d\}$, $DV_c(0)$ is chosen as the disparity vector DV_c of the current block c , and disparity estimation of the block c is over; otherwise, disparity vector DV_c have to be searched within a small range indicated by $DV_c(0)$.

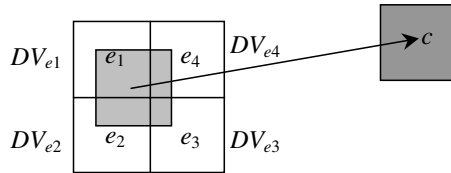


Fig. 4. Disparity vector prediction of the block c with the disparities of the block e

If the current block c does not have intra-view similar block, the Hardamrd similarity coefficient R_j between the current block c and its block e in the adjacent view is calculated. Let R_t be a threshold, if $R_j < R_t$, the block e is considered to be the matching one of the block c . Thus, disparity interpolation is used to obtain the relationship between the block c and block e . Since the block e may be composed of four sub-blocks e_1 , e_2 , e_3 and e_4 , as shown in Fig.4, the disparity vector DV_e of the block e is calculated by $DV_e = (e_1 \cdot DV_{e_1} + e_2 \cdot DV_{e_2} + e_3 \cdot DV_{e_3} + e_4 \cdot DV_{e_4}) / 64$. Then, the estimated disparity vectors $DV_{cl}(0)$ and $DV_{cr}(0)$ of the block c corresponding to the leftmost or rightmost reference view are interpolated with DV_e , $DV_{L \rightarrow R}$, and $DV_{R \rightarrow L}$. If the current view is near to the leftmost view, set the estimated disparity $DV_c(0) = DV_{cl}(0)$; otherwise, set $DV_c(0) = DV_{cr}(0)$. If $|DV_{cl}(0) - DV_{cr}(0)| < 1$, $DV_c(0)$ is considered as the disparity vector DV_c , and terminate disparity estimation of the block c ; otherwise, DV_c is estimated within a small range indicated by $DV_c(0)$.

If $R_j \geq R_t$, it indicates that the current block c has no inter-view similar block, so the best matching block in the leftmost and rightmost reference views have to be searched according to $DV_{cl}(0)$ and $DV_{cr}(0)$, and the best disparity vector among them is chosen as the disparity vector DV_c of the block c .

To reduce computational complexity further, fast searching strategies and SAD threshold are used to terminate the searching in advance. If the search is unavoidable, SADs with respect to nearby points of the point indicated by $DV_c(0)$ are firstly calculated to determine the master and slave search directions. For the master search direction, 1-D diamond search is implemented, in which 2-pixel interval is used to find the matching position cursorily, and then the best matching point is searched just around the cursorily matched position; For the slave searching direction, bi-directional search is used. The final matching point is the one with the minimum SAD selected from the master or slave search direction.

To speed up the searching, a SAD threshold, SAD_T , is used as search stop criteria. Here, the threshold SAD_T is defined by $SAD_T = (1 - R_j) SAD_{neigh}$, where R_j is the

Hadamard similarity coefficient of two blocks, SAD_{neigh} is the SAD value of the neighboring block. In searching process, if the SAD value of a point is small than SAD_T , the point is considered to be good enough so it is chosen as the best matching point and the searching process is terminated.

The proposed multi-view disparity estimation algorithm is concluded as follows

- Step-1: Let the leftmost or rightmost view as the reference view, respectively, calculate disparity map $\{DV_{L \rightarrow R}\}$ and $\{DV_{R \rightarrow L}\}$ of the two reference views.
- Step-2: Calculate Hadamard similarity coefficients between the current block and its three intra-view neighboring blocks $\{R_{na}, R_{nb}, R_{nd}\}$ according to formula (1), and let $R_n = \min(R_{na}, R_{nb}, R_{nd})$. If $R_n < R_T$, go to Step-3, otherwise go to Step-4.
- Step-3: Use the disparity vector and reference view of neighboring block as estimated disparity vector $DV_c(0)$ and reference view of the current block. Compare $SAD_c(0)$ obtained with $DV_c(0)$ to SAD_a , SAD_b and SAD_d of the blocks a , b and d , if $SAD_c(0) \leq \text{med}\{SAD_a, SAD_b, SAD_d\}$, $DV_c(0)$ is considered as the disparity vector DV_c of the block c , and then go to Step-8. Otherwise, go to Step-6.
- Step-4: Calculate Hadamard similarity coefficient R_j between the current block c and block e according to formula (2), if $R_j < R_t$, go to step 5, otherwise go to Step-7.
- Step-5: Obtain $DV_{cl}(0)$ and $DV_{cr}(0)$ of the block c corresponding to the leftmost and rightmost reference views by interpolating with the disparity map $\{DV_{L \rightarrow R}\}$ and $\{DV_{R \rightarrow L}\}$. Choose $DV_{cl}(0)$ or $DV_{cr}(0)$ as $DV_c(0)$ in terms of the distance between the current view and the leftmost or the rightmost view. If $|DV_{cl}(0) - DV_{cr}(0)| < 1$, $DV_c(0)$ is regarded as the disparity vector DV_c , and then go to Step-8, otherwise go to Step-6.
- Step-6: Do single reference disparity estimation by using $DV_c(0)$ as the initial disparity vector of the block c , then go to Step-8.
- Step-7: Do two reference view disparity estimation by using $DV_{cl}(0)$ and $DV_{cr}(0)$ as the initial disparity vectors, the disparity vector with the minimum SAD is chosen as final disparity vector of the block c .
- Step-8: Finish the search of the block c , and then turn to estimate the disparity of the next block.

4 Experiments and Discussions

In order to evaluate the proposed multi-view disparity estimation algorithm, experiments have been performed on two test sequences of real data called “Xmas”, and “Note”. “Xmas” is available from Tanimoto Lab, 101 viewpoint images are captured synchronously with the camera interval of 3mm, the image resolution is 640×480 . Here, we select 10 of 101 views, with the camera interval of 30mm, as a test set. “Note” is captured by a shifted camera with the interval of 30mm. Fig.5 shows three views of the two test sequences. Fig.6 shows four blocks in one image of Xmas. From the figure, block 1 and block 2 are within edge regions, but do not have the same characteristic. Block 3 and block 4, on the other hand, are within the same smooth region, so they have the same characteristic. Hadamard similarity coefficients of the four blocks are $R_{1,2}=0.258$, $R_{1,3}=0.474$, $R_{2,3}=0.243$, and $R_{3,4} \approx 0$, which mean that

the block 1, block 2 and block 3 are different with each other, but block 3 is quite similar to block 4 because Hadamard similarity coefficient is close to 0. Thus, Hadamard similarity coefficient is able to describe the similarity of two blocks. Moreover, its computational complexity is very low.



(a) Three views of Xmas multi-view images



(b) Three views of Note multi-view images

Fig. 5. Two test sets of multi-view images

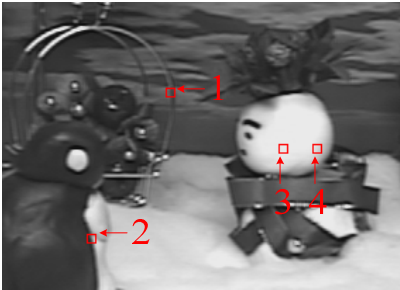


Fig. 6. Four blocks in Xmas test data

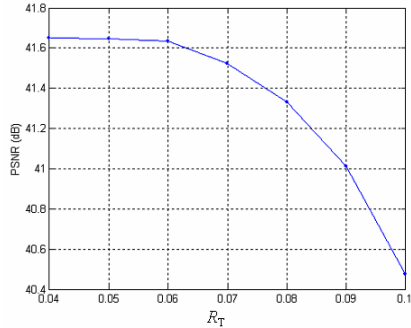


Fig. 7. Coding results with different R_T

Fig.7 shows PSNRs of the decoded signals with respect to different threshold R_T . Smaller R_T will increase high computational complexity, so here R_T is set to be 0.06. Fig.8 gives rate-distortion curves with respect to full search algorithm (FS), direct limit fast search algorithm (DLS)^[8] and the proposed algorithm. Experimental results show that they achieve almost the same rate-distortion performance.

Table 1 compares the computational complexity of the three algorithms. Compared with the FS algorithm, the proposed algorithm saves up to 98.4%~97.8% coding time. The proposed algorithm can also obviously reduce the computational complexity even compared with other fast search algorithm, such as DLS algorithm. Table 2 lists the number of searching points with respect to the three algorithms, from which similar conclusion can be made as from Table 1.

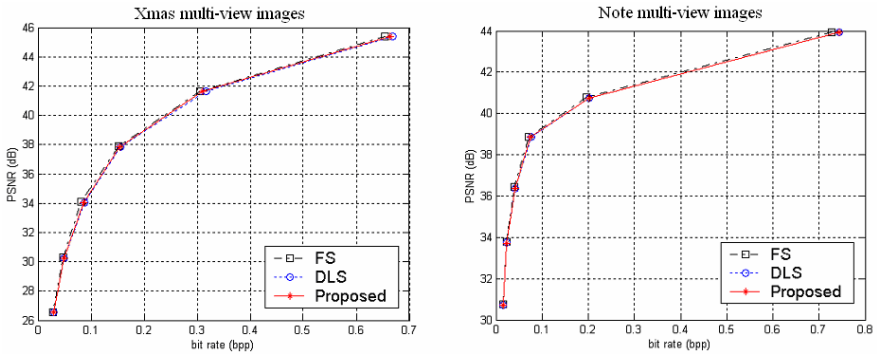


Fig. 8. (a) Experimental results of Xmas

(b) Experimental results of Note

Table 1. Comparison of three algorithms with full coding processing

	Xmas			Note		
	bits	PSNR	time	bits	PSNR	time
FS	937904	41.66	100%	604792	40.80	100%
DLS	956236	41.64	16%	620369	40.73	13.5%
Proposed	954267	41.64	2.2%	619481	40.73	1.6%

Table 2. Comparison of three algorithms only for search processing

Test data	algorithm	PSNR	number of search point	speedup
Xmas	FS	41.66	768	1
	DLS	41.64	105.2	7.3
	Proposed	41.64	12.2	62.9
Note	FS	40.80	768	1
	DLS	40.73	80.4	9.6
	Proposed	40.73	9.4	81.7

5 Conclusion

Disparity estimation is a key technique in multi-view data compression, since it is an efficient tool to eliminate the redundancy among different views. In this paper, a novel fast multi-view disparity estimation algorithm based on Hadamard similarity coefficient is proposed by using prediction of initial search point, selection of reference view, determination of the best disparity vector and strategies of search stop, in order to reduce the computational complexity of disparity estimation. Experimental

results show that the proposed algorithm can significantly reduce the computational complexity in the multi-view video coding at the cost of tiny descent of rate-distortion performance compared with the full search algorithm.

Acknowledgment

This work was supported by the Natural Science Foundation of China (grant 60472100), the Natural Science Foundation of Zhejiang Province (grant RC01057, 601017, Y105577), the Ningbo Science and Technology Project of China (grant 2003A61001, 2004A610001, 2004A630002), and the Zhejiang Science and Technology Project (Grant 2004C31105).

References

1. ISO/IEC JTC1/SC29/WG11 N5877: Application and Requirements for 3DAV. Trondheim, (2003)
2. Jiang, G., Shao F., Yu, M., et al.: Efficient Block Matching for Ray-Space Predictive Coding in Free-Viewpoint Television Systems, Lecture Notes in Computer Science, Vol. 3980. Springer-Verlag, Berlin Heidelberg New York (2006) 307-316
3. Jiang, G., Fan, L., Yu, M., et al.: An Approach to Fast Ray-space Interpolation for Free Viewpoint Video System, Lecture Notes in Artificial Intelligence, Vol.3802. Springer-Verlag, Berlin Heidelberg New York (2005) 935-940
4. ISO/IEC JTC 1/SC 29/WG 11N7328:Introduction to Multi-view Video Coding, Poznan, Poland, (2005)
5. Grammalidis, N., Strintzis, M.: Disparity and occlusion estimation in multi-ocular systems and their coding for the communication of multi-view image sequences. IEEE Trans. on CSVT, (1998) 328-344
6. Kimata, H., Kitahara, M., Kamikura, K., et al.: Multi-view video coding using reference picture selection for free viewpoint video communication, Proc. of Picture Coding Symposium, (2004), 499-502
7. Lopez, J., Kim, J.: Block-based illumination compensation and search techniques for multi-view video coding, Proc. of Picture Coding Symposium, (2004), 509-514
8. Woo, W., Ortoga, A.: Stereo images compression with disparity compensation using the MRF mode, Proc. of The SPIE, Vol.2727, (1998), 28-39

AddCanny: Edge Detector for Video Processing

Luis Antón-Canalís¹, Mario Hernández-Tejera¹, and Elena Sánchez-Nielsen²

¹ Institute for Intelligent Systems and Numerical Applications in Engineering - IUSIANI. University of Las Palmas de Gran Canaria (ULPGC). Campus Universitario de Tafira, Las Palmas, Spain
{lanton, mhernandez}@iusiani.ulpgc.es

² Departamento de Estadística, Investigación Operativa y Computación, 38271 University of La Laguna, S/C Tenerife, Spain

Abstract. In this paper, we present AddCanny, an Anisotropic Diffusion and Dynamic reformulation of the Canny edge detector. The proposal provides two modifications to classical Canny detector. The first one consists of using an anisotropic diffusion filter instead of a Gaussian filter as Canny does in order to obtain better edge detection and location. The second one is the replacement of the hysteresis step by a dynamic threshold process, in order to reduce blinking effect of edges during successive frames and, therefore, generate more stable edges in sequences. Also, a new performance measurement based on the Euclidean Distance Transform to evaluate the consistency of computed edges is proposed. The paper includes experimental evaluations with different video streams that illustrate the advantages of AddCanny compared to classical Canny detector.

1 Introduction

Edge detection plays an important role in image processing domain, vision system performance and different applications of computer vision such as object detection, tracking or recognition problems. Most edge detection methods work on the assumption that an edge occurs where there is a discontinuity in the intensity function or a steep intensity gradient in the image. Different operators have been proposed using first order differential methods for edge detection such as Robert's Cross, Prewitt or Sobel [2]. Currently, the most common approach widely used is the Canny edge detector [4], which is aimed as the optimal edge detector. Formally, this approach specifies an objective function to be optimized, with the following optimization constraints: (i) maximize the signal to noise ratio to give good detection, (ii) achieve good localization to accurately mark edges and (iii) minimize the number of responses for a single edge.

Traditionally, Canny edge detector has been successfully applied in still image based applications. There are works which improve Canny's performance [6] [5] and also those which propose a replacement with different gradient based edge detectors [1]. Dynamic hysteresis is studied in [11] although not used in a Canny detector. However, video processing using Canny edge detector introduces new issues to be addressed:

- Consistency: video processing applications using edge detection methods require edge based features to be detected consistently and efficiently at the spatiotemporal domain. However, the use of Canny detector, which has been mainly designed for still images, introduces instability situations in detected edges during the sequence. These can be noticed as blinking edges, that is, edge features that appear in one frame but suddenly disappear in the next one. This blinking effect takes place when there are certain changes between consecutive images in a stream, mainly originated by sensor noise effects or small lighting changes. As a result, under continuously operating conditions, detection quality cannot be minimally ensured for video processing.
- Complexity: computational complexity in real-time video processing is a crucial factor. Therefore, operations to be computed must be optimized.
- Threshold computation: gradient-based edge methods require threshold values to be fixed or chosen manually for certain lighting conditions. In order to provide stable edge features in video streams, the adaptation of these threshold values must be automated.
- Performance measurement: measuring edge detector performance in video streams is a complex task. It is difficult since edges are subjective features and thus difficult to define.

In this paper, we propose a new approach to compute consistency of edge detection on video streams. Our approach is focused on Canny edge detector with the purpose of preserving good detection, localization and minimal response. We improve the edge detector method, obtaining a higher edge consistency in video streams, focusing on the reduction of parameters in order to achieve a general purpose edge detector. We reduce the blinking effect of edges by the substitution of the hysteresis step by a dynamic threshold process, aided by an anisotropic diffusion filter. In order to evaluate the edge detector performance, we propose a new edge consistency measurement based on the Euclidean Distance Transform [7]. The structure of this paper is as follows: the Canny detector, including an analysis of how the hysteresis step is the main responsible for blinking edges in video sequences, is described in Section 2. Section 3 describes the effect of hysteresis in video sequences. Section 4 explains our proposal and its advantages, which are proved experimentally in Section 5, where a measurement performance for edge consistency between frames is proposed and described. Conclusions and future work are presented in section 6.

2 Canny Edge Detector

Canny edge detector [4] is a three steps procedure that leads to the three desirable features of an optimal edge detector: good detection - the algorithm should mark as many real edges in the image as possible; good localization - found edges should be as close as possible to edges in the real image; minimal response - a given edge in the image should be marked only once, and where possible, image noise should not create false edges. Canny's steps are described briefly in the following paragraphs:

Step 1: Smoothing

A 2D Gaussian mask is convolved with the image in order to remove high frequency noise. The mask aperture can be selected to perform a stronger smoothing effect, but it may produce edge delocalization.

Step 2: Non-maximal suppression

Image’s horizontal and vertical partial derivatives are obtained and used to calculate gradient magnitude and direction. Using them, it is possible to eliminate (set to zero) those pixels which gradient magnitude is not local maxima along the direction that is perpendicular to the edge, finally obtaining a thin-gradient image like the one shown in Fig.4c.

Step 3: Hysteresis

Hysteresis is a powerful mechanism in many processes. The main idea is that an element becomes more important for a certain process if an already important element exists nearby. In the Canny edge detector, hysteresis is a recursive procedure that involves two thresholds: l and h (low and high). According to them, for each pixel in the thin-gradient image, it is marked as an edge pixel if its magnitude value is higher than h , and rejected if it is lower than l . For edge pixels, their neighbor pixels are considered, and marked as edge pixels (and thus recursively analyzed afterwards) if their magnitude value lies between l and h .

3 Hysteresis in Video Sequences

Thanks to the hysteresis process, edges become complete or not broken in still images. In video sequences, however, hysteresis produces undesired effects related to how thresholds affect the process. First of all, every edge created from a single pixel higher than the higher threshold h on a certain frame, may completely disappear in the following frame if that pixel’s value descends below h . Suppose the following 1D example, after Gaussian convolution and non-maxima suppression, where upper rows contains values to which hysteresis with $l = 4$ and $h = 8$ is applied:

Frame n	1 2 7 8 8 8 8 8 8 8 9 8 8 8 8 8 8 7 2 1
Edge	0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0
Frame n+1	1 2 3 8 8 8 8 8 8 8 8 8 8 8 8 8 8 3 2 1
Edge	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0

While on frame n a whole edge is found thanks to a single pixel, on frame $n + 1$ it disappears because the same pixel is not above h . A second situation shows how an edge is shortened because of a similar reason:

Frame n	1 2 7 9 7 6 5 7 8 8 8 8 8 8 8 8 7 2 1 4
Edge	0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 0
Frame n+1	1 2 3 9 7 6 4 7 8 8 8 8 8 8 8 8 8 3 2 1
Edge	0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0

A third effect, less probable but still possible, involves an edge which is broken in two smaller edges if two pixels' values descent below the lower threshold:

Frame n	498758888888888857894
Edge	01111111111111111110
Frame n+1	198748888888888847891
Edge	01110000000000001110

Full edges are lost or broken even when their pixel's values are close enough to h . This way, Canny edge detector produces complete edges, visually attractive, but it may lose information in video sequences. Due to sensor noise and small light changes these events constantly appear in video sequences, and it is easy to visually observe them, as seen in Fig.1. This variation may be reduced choosing the right h and l parameters. However, different sequences will require different thresholds, and the blinking effect does not disappear due to how hysteresis works. Moreover, if a sequence includes light variations, the two thresholds may have to be reconfigured during the sequence. If thresholds are not chosen properly, edges shall not be successfully detected and the blinking effect may be dramatically increased, as shown in Fig.1.

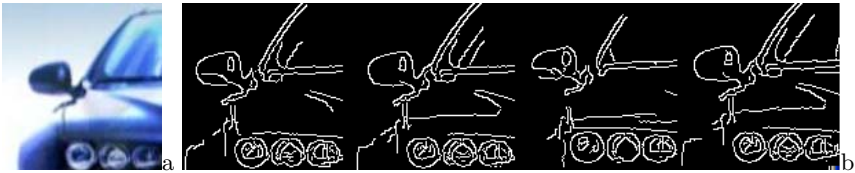


Fig. 1. Blinking edges example in Car sequence: From a sequence with a smooth light transition from a lighter image to the one shown on the left, four frames from a Canny with $l=100$ and $h=200$) are shown. Notice the rear-view mirror and how edges vary abruptly within continuous frames.

4 An Edge Detection Proposal for Video Streams: AddCanny

AddCanny, which stands for Anisotropic Diffusion Dynamic Canny, is an approach based on two main changes to original Canny's proposal: the substitution of the Gaussian filter step with an anisotropic diffusion filter [8] and the replacement of the hysteresis step with a dynamic thresholding operation. Thus, edge localization and minimal response of a Canny edge detector are conserved, but a higher edge consistency in video sequences is obtained. During the whole process, we focus on the reduction of the amount of parameters in order to achieve a general purpose edge detector.

4.1 Anisotropic Diffusion

In our proposed solution, the Gaussian convolution in the noise suppression step is replaced with a strong anisotropic diffusion filtering process. Homogeneous regions are strongly smoothed and noise and weak edges are removed while sharp boundaries are preserved. Anisotropic filtering algorithms remove noise from images while preserving object boundaries. Tomasi and Manduchi's Bilateral filter [9] and Perona and Malik's anisotropic diffusion [8] achieve this through intensity/color and gradient dependent computations between pixels. While Bilateral Filter seems to produce better results (see Fig.2c), it is not suitable for real time processing. A low number of Anisotropic Diffusion iterations are much faster and images are properly filtered for our purposes as seen in Fig.2d. Perona and Malik in [8] analyzed filtering processes via partial differen-

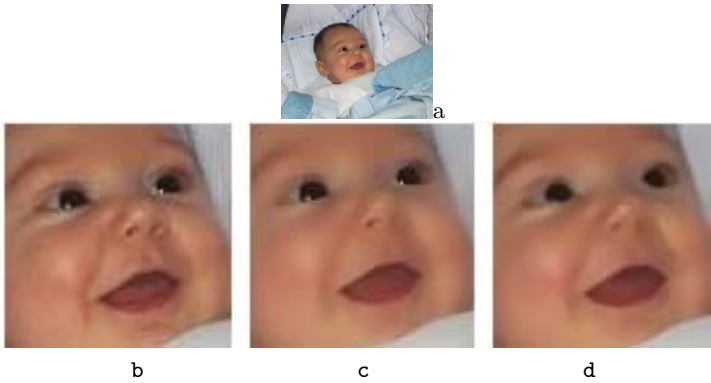


Fig. 2. a) Jorge image, 320x240 pixels RGB, unfiltered. b) Original image detail c) Bilateral Filter result (range 3, domain 5, aperture 11): 406ms for the whole image d) Simplified Anisotropic Diffusion filter result: 3 iterations : 140ms for the whole image. (Xeon CPU, 3Ghz).

tial equations (PDE) and proposed a nonlinear modification of images replacing the classical isotropic diffusion equation with the following equation:

$$\frac{\partial I(x, y, t)}{\partial t} = \text{div}[g(\|\nabla I\|)\nabla I] \quad (1)$$

where ∇I is the image gradient magnitude and $g(\|\nabla I\|)$ is an *edge-stopping* function, chosen in order to satisfy $g(x) \rightarrow 0$, when $x \rightarrow \infty$, so that diffusion is performed nonlinearly, applied in non-edge zones and *stopped* across edges. The following function was originally suggested by Perona and Malik [8]:

$$G_1(x) = \frac{1}{(1 + \frac{x^2}{K^2})} \quad (2)$$

For a review of more stopping functions, see [3]. In order to maximize the diffusion process, we propose the following modification of a SigmoidS [10] for the $g(\cdot)$ function, avoiding the positive constant K parameter in 2.

$$G_2(x) = \begin{cases} 1 & x \leq 0 \\ 1 - x^2 & 0 < x \leq 0.25 \\ 0 & x > 0.25 \end{cases} \quad (3)$$

Ranges used in proposed function 3 are based on a visual analysis of the image gradient ∇I after a dynamic range amplification, setting gradient values in $[0, 1]$. It is reasonable to agree that, with values below 0.25, edges begin to be considered weak enough to allow diffusion, as can be observed in terms of grey values in Fig.3. This way, anisotropic diffusion becomes strong enough and only a few

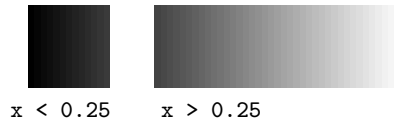


Fig. 3. Gray-scale values, representing edge strengths in a gradient image ∇I

iterations are needed in order to smooth homogeneous regions (three in our tests).

4.2 Thin-Gradient Image Information

Given a frame f from a video sequence, gradient magnitude and direction are obtained using two partial derivative Sobel kernels on the filtered image $\alpha(f)$, and pixels that are not local maxima in gradient magnitude along the direction that is perpendicular to the gradient direction are suppressed. At this point we have a grayscale image $\beta(f)$ obtained from a filtered image $\alpha(f)$ that contains thin and localized edges, as good as Canny’s (because we have followed the same basic steps), with useful information about edge strengths. It is noticeable how edge images are usually considered as binary images, even when non-binary images contain much more information in relation to edge strength (see Fig.4c). However, if a binary edge image is needed, the hysteresis step in the Canny detector is usually applied, obtaining Fig.4d.

4.3 Hysteresis Replacement

As seen in Section 3, it is the hysteresis step what seems to produce blinking edges in video sequences. In order to prove this observation, we propose a replacement of the hysteresis step with a simple image binarization using a dynamic threshold on each frame from a video sequence. Given the thin-gradient image $\beta(f)$ obtained from the filtered version $\alpha(f)$ of a certain frame f , a threshold T_f is adjusted automatically as the mean of those pixel values p_i in $\beta(f)$ above zero:

$$T_f = s \cdot \frac{\sum p_i \forall p_i \in \beta(f), p_i > 0}{N} \quad (4)$$

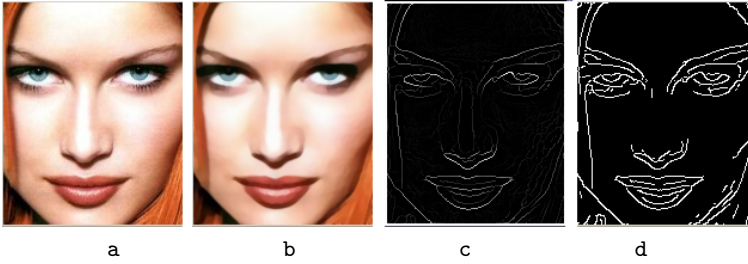


Fig. 4. Laetitia image, a) unfiltered f , b) Anisotropic Diffusion $\alpha(f)$, c) greyscale thin-gradient image $\beta(f)$, d) OpenCv's Canny $l=64$, $h=128$ binary image

where N is the number of non-zero values in $\beta(f)$ and s is a weighting factor. This particular simple threshold operation does not return complete edges like a more complex hysteresis process would do, but it returns edges good enough. Moreover, T_f is a dynamic parameter which adapts to each frame's lighting conditions, so it becomes useful in sequences with environmental variations. However, edges are subjective features from images, and usually depend on what the user wants to find in a given image. In order to allow a certain output control, a single weighting factor s is introduced, which stands for *edge strength*. Threshold T_f is weighted using s , which value may be chosen around 1.0 (below 1.0 would mean more edges, because the threshold is lowered, while above 1.0 means more edges), so it is possible to get more or less edges depending on user's preferences or application needs. In our experiments, we set $s = 1.0$ when possible.

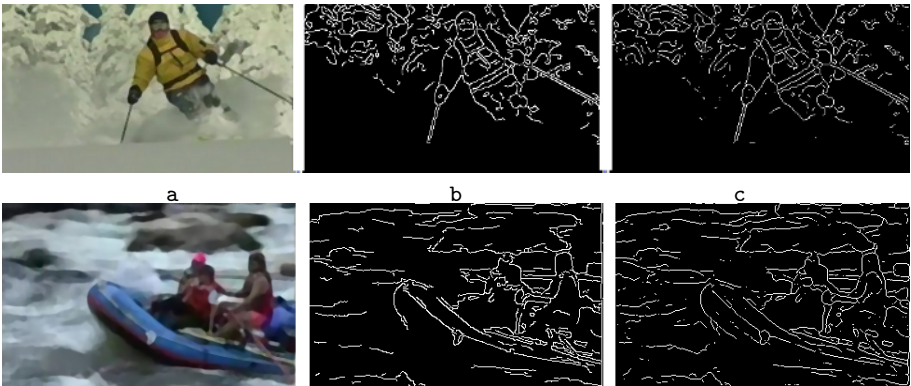


Fig. 5. Two examples of our approach compared to Canny edge detector a) Original image, b) Canny (top) (70, 120), Canny (bottom) $l=90$, $h=140$, c) Proposed approach ($s = 1.0$)

4.4 Anisotropic Diffusion Effect in Presence of Strong Movement

In certain video sequences captured in the presence of strong movements, both egomotion or moving objects in the scene, there may be a certain amount of blurriness in images, caused when the camera recording frequency is not fast enough to properly capture fast motion. The anisotropic diffusion filter adds definition to these images, allowing the detection of borders with a higher quality, recovering edges where Gaussian filters smooths them, as seen in Fig.6. Anisotropic



Fig. 6. A frame from Elena sequence: a) Original Image, b) Filtered Image, c) Canny $l=170$, $h=220$, d) Our approach ($s = 1.0$)

diffusion in moving sequence aids in defining moving objects' boundaries, as seen on the background wall in Fig.6. Canny may manage these images, but its parameters may have to be tuned to be sensitive enough, which also adds unwanted edges.

5 Empirical Measurement of Edge Stability in Video Streams

In order to test edge consistency throughout consecutive frames, six sequences are used as experimental set in this paper. They show a varied amount of situations, both indoor and outdoor sequences, with and without moving objects and varying light conditions. For a random frame in each sequence, we adjusted original Canny's h and l parameters, always using a 3×3 Gaussian filter, in order to get good edges. Then we tuned our method with three anisotropic diffusion iterations and we set the weighting factor s in order to compute an amount of edges similar to Canny's, which we usually obtained using $s = 1.0$. Finally, we processed all the sequences and measured stability. The amount of edge variation was measured computing the distance between a frame from a given sequence and the next one using Distance Transform (DT) images [7]. In a DT image, every non-object pixel from the source binary image is labeled with its Euclidean Distance to the closest edge pixel in the source image, as seen in the following image: The subtraction of DT images allows the computation of a distance between binary images, measuring not only coincidences, as a subtraction of plain binary images would return, but distance between edges present in both images.

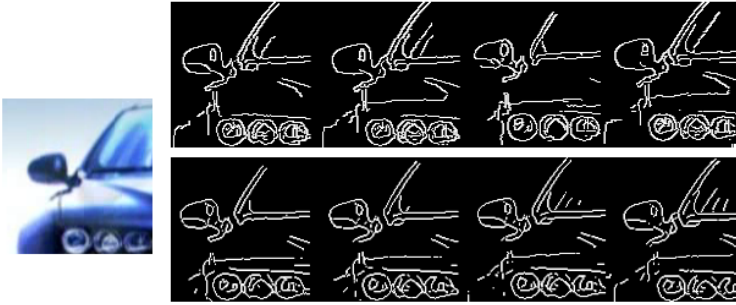


Fig. 7. Results from AddCanny applied to the same Car sequence from Fig. 1. Canny ($l=100$, $h=200$) is shown on top row, AddCanny($s = 1.0$) in bottom row. Some new details appear as light changes, but changes are much smoother between consecutive frames with our approach.

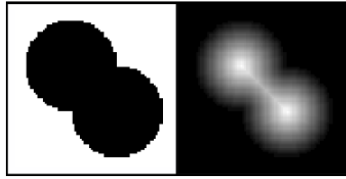


Fig. 8. Binary image and its DT image. Pixels are labelled with their distance to the closest object pixel (white pixels).

For each frame, a binary image was obtained using Canny and our approach. Then, distances between binary images were computed using Distance Transform (DT) images. This way, the sum of squares of absolute differences between pixels from two DT images shows a measurement of the amount of variation between both of them. Given the DT images of two binary images img_1 and img_2 with the same size, their difference is measured as shown in 5.

$$d(img_1, img_2) = \sum (||DT(img_1)_i - DT(img_2)_i||^2) \forall i \in img_1 \quad (5)$$

We use the squared difference to give more importance to higher values, because most edges are properly localized in both methods and thus their majority would affect and hide the negative effect of blinking boundaries. Fig.9 shows normalized results for the six sequences. On most cases our approach achieves more consistent edges, except for the Office sequence, where results are slightly worse. This is a particular sequence because there are no moving objects. In this case, even though Canny and AddCanny produce visually similar edges, Canny's unbroken edges achieve better measurement values according to 5, mainly due to the presence of many independent edge pixels created by AddCanny. Note that each sequence needs a different pair of Canny parameters, while only the Duck and Office sequence needed a different s parameter in order to raise ($s = 0.5$,

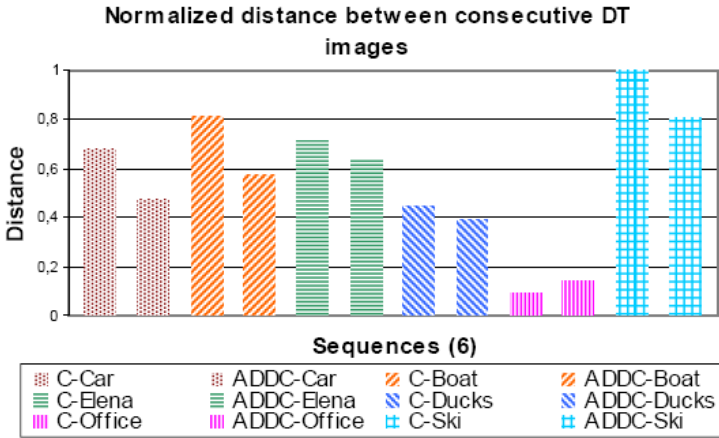


Fig. 9. Edge consistency test for six sequences (each pattern one sequence). For each sequence, the first column represents results with original Canny and the second column results for AddCanny.

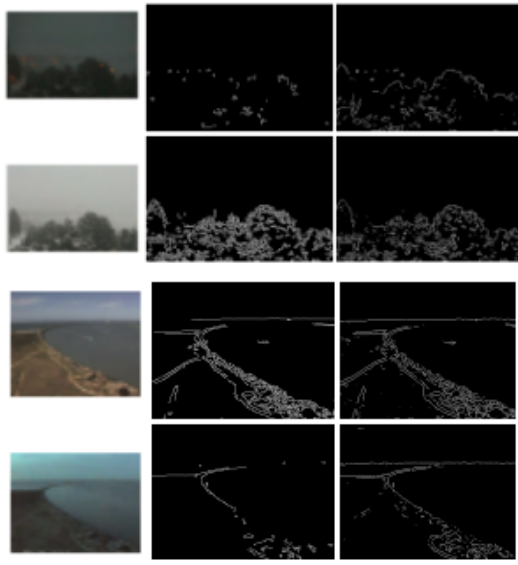


Fig. 10. Images took from online web cams at different daytimes. Original images are shown on left row. Canny $l=64$, $h=128$ is shown on center row, and our approach ($s=1.0$) on right row.

Office) or decrease ($s = 1.5$, Ducks) the number of edges in order to return an amount of edge pixels similar to Canny's. In those sequences that cover a wide spectrum of light conditions, like certain security cameras which record during

day and night cycles, Canny's original proposal is not able to perform properly because of its static parameters. In those cases, our approach behaves much better due to the dynamic nature of its threshold parameter. It is difficult or plainly impossible to obtain a pair of parameters for Canny method than could manage these kind of sequences, this situation forces the use of dynamic thresholds.

6 Conclusions

In this paper, we have presented an edge detector based on Canny's proposal for edge detection, preserving good detection, localization and minimal response, but reducing parameters and aiming towards edge consistency in video streams. AddCanny reduces edge inconsistency between continuous frames, avoiding blinking edges like those that may appear using classical Canny edge detector. The substitution of the Gaussian filter with an anisotropic diffusion filter reduces noise while preserving objects boundaries accurately. Furthermore, it sharpens blurred images caused by strong movements, avoiding the disappearance of edges in those situations. AddCanny adjusts automatically its operation parameter, regulating its sensitivity depending on image conditions. In order to compare edge consistency in video streams, a stability measure based on the Distance Transform image has been proposed. Our procedure has been presented comparing it to the use of classical hand-tuned Canny approach and results conclude that AddCanny shows a better performance than classical Canny for video stream edge detection purposes. Moreover, it sustains the quality of detection when lighting conditions change. As future work, the anisotropic diffusion filter will be extended in order to include temporal information, local thresholding will be used instead of the current global threshold T_f , and a modification of the stability measure will be studied in order to give the longest contours a higher weight, thus correcting the current effect of noisy and small contours in its value. Finally, due to the nature of thresholds in image processing, we also consider the application of fuzzy logic to the binarization step.

References

1. A. Averbuch, B. Epstein, N. Rabin, E. Turkel. Edge-Enhancement Postprocessing Using Artificial Dissipation *IEEE Transactions on Image Processing* Volume 15, Issue 6, 1486–1498 June 2006.
2. B. Jahne. Digital Image Processing. *Springer 4th edition*, 1997. ISBN: 3540240357
3. M. J. Black, G. Sapiro, D. Marimont, D. Heeger Robust anisotropic diffusion. *IEEE Transactions on Image Processing, Special issue on Partial Differential Equations and Geometry Driven Diffusion in Image Processing and Analysis*, 7(3), 421–432, 1998.
4. J. Canny. A Computational approach to Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 8(6), 1986.
5. D. Demigny. On optimal linear filtering for edge detection *IEEE Transactions on Image Processing* Volume 11, Issue 7, 728–737, July 2002.

6. Gang Liu and R.M. Haralick. Two practical issues in Canny's edge detector implementation *Proceedings of 15th International Conference on Pattern Recognition*. Volume 3, 3-7, 676–678 Sept. 2000.
7. D. W. Paglieroni. Distance Transforms. *Computer Vision, Graphics and Image Processing : Graphical Models and Image Processing*, 54, 56–74, 1992.
8. P. Perona and J. Malik. Scale-space and edge detection using anisotropic diffusion. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 12(7), 629–239, 1990.
9. C. Tomasi and R. Manduchi. Bilateral filtering for gray and color images. *Sixth International Conference on Computer Vision*, 839–846. New Delhi, India, 1998.
10. L. A. Zadeh et. al. Fuzzy sets and their applications to cognitive and decision processes *Academic Press*, 1975. ISBN: 0127752609
11. Junfeng Ge, Yupin Luo and Deyun Xiao. Adaptive hysteresis thresholding based pedestrian detection in nighttime using a normal camera *IEEE International Conference on Vehicular Electronics and Safety.*, 46–51. 14-16 Oct. 2005.

Video-Based Facial Expression Hallucination: A Two-Level Hierarchical Fusion Approach*

Jian Zhang, Yueting Zhuang**, and Fei Wu

College of Computer Science & Technology, Zhejiang University, Hangzhou 310027,
P.R. China
zhangsdust@yahoo.com.cn, {yzhuang, wufei}@cs.zju.edu.cn

Abstract. Facial expression hallucination is an important approach to facial expression synthesis. Existing works mainly focused on synthesizing a static facial expression image given one face image with neutral expression. In this paper, we propose a novel two-level hierarchical fusion approach to hallucinate dynamic expression video sequences when given only one neutral expression face image. By fusion of local linear and global nonlinear subspace learning, the two-level approach provides a sound solution to organizing the complex video sample space. Experiments show that our approach generates reasonable facial expression sequences both in temporal domain and spatial domain with less artifact compared with existing works.

1 Introduction

Facial expression synthesis techniques have been widely used in the fields of human-computer interaction, film making and game entertainment. However, due to the sensitivity to human face, facial expression synthesis is a challenging topic in computer vision. Qingshan Zhang et al. [1] developed a geometry-driven facial expression synthesis system which could generate photorealistic facial expressions through blending sub-region texture images according to the facial feature positions. Liu et al. [2] proposed an expression mapping approach based on expression ratio image (ERI). Combining illumination changes of one person's expression with geometry warping, they mapped an ERI to arbitrary face and generated more expressive facial expressions. Besides the analogy and retargeting approaches mentioned above, learning based techniques for face synthesis have also been explored by many researchers. "Face hallucination" was first proposed by Simon Baker et al. [3], the motivation was to create high resolution version of an input low resolution face image by sample learning. Based on a complicated probabilistic model, Liu et al. [4] built a two step approach to hallucinate human faces, the global parametric model aimed at recovering global face image while the local non-parametric model contributed to generating face

* This work is supported by the key program of National Natural Science Foundation of China (No.60533090), National Science Fund for Distinguished Young Scholars (No.60525108), 973 Program (No.2002CB312101), Science and Technology Project of Zhejiang Province (2005C13032, 2005C11001-05).

** Contact author.

details. Similar architecture was adopted by [5, 6, 7, 8] for face hallucination. In [9], the authors developed an eigen-transformation approach, through which high resolution face image was synthesized by weighted linear combination of the training samples. In [10], the authors extended face hallucination to synthesize facial expressions. Given face image with neutral expression, they created reasonable facial expression from sample images.

In this paper, we propose a facial expression synthesis approach based on face hallucination, compared with [10], the novelty of our work mainly lies in two points: first, instead of synthesizing single facial expression image, we propose a video based expression hallucination approach which generates an expression video sequence given only one image; second, a hierarchical framework is adopted to perform the two-level subspace learning on video samples, this approach is a fusion of linear and nonlinear subspace learning. Our approach is based on neighbor reconstruction, two main differences make our approach work better than the weighted linear combination method adopted in [9]: first, applying manifold learning algorithm to select nearest reconstruction neighbors instead of using the whole sample space to compute the linear combination weights; second, radial basis function regression rather than weighted linear combination is adopted to gain final expression video sequence.

The paper is organized as follows: in section 2, the framework of a two-level hierarchical fusion approach for video expression hallucination is introduced; in section 3, the hallucination procedure is presented in detail; experiments and discussion are given in section 4; section 5 concludes this paper.

2 Framework of the Two-Level Hierarchical Fusion

The training samples comprise tens of video clips, each video clip represents one kind of facial expression of one specific person from neutral to apex. Our goal is to hallucinate expression video sequences of a new test subject based on sample videos, the input is a single frontal face image with neutral expression.

Due to the high dimensionality of video samples, organizing the sample space well becomes a challenging problem. Here, we propose a hierarchical approach to perform training and synthesizing, which includes two levels: local linear subspace learning and global nonlinear subspace learning.

2.1 Local Linear Subspace Learning

In this level, each training sample (a video sequence) is considered to construct a local linear subspace. Principal component analysis (PCA) [11] has been proved effective in learning such a linear subspace. So, in this level, we use PCA to compute eigen-representation of each sample offline for further use. Given a video sequence, we stack columns of each frame into one vector and integrate all the vectors to form a sample matrix X , and then we compute $\tilde{X} = (X - \bar{X}) / N^{1/2}$ to register X where \bar{X} is the mean value and N is the number of samples. To deal with the problem caused by high dimensionality, we perform QR factorization to gain $[q, r] = QR(\tilde{X})$, then SVD decomposition is imposed on r to get $[u, s, v] = SVD(r)$, and then eigenvectors can be obtained by

$U = q \cdot u$, we do so for solving the problem in a numerically stable way. Thus, we can project any frame f on U to get the reconstruction coefficients $y = U^T \cdot (f - \bar{X})$, and f can be reconstructed by $\tilde{f} = U \cdot y + \bar{X}$. So, we store each sample's eigenvectors U , coefficients y and mean value \bar{X} as the eigen-representation for expression synthesis.

2.2 Global Nonlinear Subspace Learning

In global nonlinear subspace learning, given an input face image with neutral expression as a high dimensional data point, we aim at finding its nearest neighbors among the first frames (neutral) of the video samples by Locally Learning Embedding (LLE) algorithm [12]. LLE is an unsupervised manifold learning algorithm that computes low dimensional, neighborhood preserving embeddings of high dimensional input and recovers the global nonlinear structure from locally linear fits. According to LLE, each high dimensional data point can be reconstructed by weighted linear combination of its neighbors. The reconstruction weights reflect intrinsic geometric properties of the data that are invariant when high dimensional data points are transformed to low dimensional coordinates. The process of LLE algorithm is briefly described as below:

Step 1. Selecting K closest neighbors for each point using a distance measure such as the Euclidean distance.

Step 2. Solving for the manifold reconstruction weights [12]. The reconstruction errors are measured by the cost function:

$$\mathcal{E}(w) = \sum_{i=1}^N \|X_i - \sum_{j=1}^N W_{ij} X_j\|^2 \quad (1)$$

where X_i is a data point and $\mathcal{E}(w)$ is the sum of the squared distances between all data points and their reconstruction neighbors. The weights W_{ij} represent the contribution of the j th data onto the i th reconstruction.

Two constraints should be obeyed:

1. $\sum W_{ij} = 1$
2. $W_{ij} = 0$ if x_j is not a neighbor of x_i

The weights are then determined by a least squares minimization of the reconstruction errors.

Step 3. Mapping each high dimensional data X_i to a low dimensional coordinate Y_i .

This is done by minimizing the cost function representing locally linear reconstruction errors:

$$\Phi(Y) = \sum_{i=1}^N \|Y_i - \sum_{j=1}^N W_{ij} Y_j\|^2 \quad (2)$$

After LLE implementation, we gain the low dimensional coordinates Y of both the neighbor samples and the input image, then expression sequence synthesis can be performed based on Y and U , y , \bar{X} which have been computed through local linear subspace learning. In deed, the two-level hierarchical approach is a fusion of nonlinear and linear subspace learning, the local level aims at simplifying the video hallucination by eigen-representation in temporal domain, the global level contributes to providing optimized expression appearance in spatial domain. The detailed expression synthesis procedure will be discussed in next section.

3 Facial Expression Sequence Hallucination

Our proposed hierarchical fusion approach performs the local linear learning offline only once, while the global nonlinear learning is performed each time when a test subject comes.

In detail, let I_{in} be the input subject image and I_{tr} be the first frames of training samples, we integrate I_{in} and I_{tr} into one matrix form (each image can be stacked into one column vector) and find the N nearest neighbors of I_{in} in I_{tr} by LLE, also, the reconstruction weights as well as the low dimensional manifold coordinates Y_{in} and Y_{tr} of these images are simultaneously computed, here Y_{in} corresponds to I_{in} and Y_{tr} corresponds to I_{tr} . The low dimensional coordinates of Y_{in} 's N nearest neighbors can be denoted as Y_{nb} . Since the eigenvectors U , coefficients y and mean value \bar{X} of each video sample have been computed through local level learning, we choose U_{nb} , y_{nb} , \bar{X}_{nb} (the eigen-representations) of those nearest neighbors and Y_{nb} as training data to synthesize the eigenvectors U_{in} , coefficients y_{in} and mean value \bar{X}_{in} of the expression sequence corresponding to the input image I_{in} .

The hallucination is well done by radial basis function (RBF) regression [13]. The RBF regression function takes the form:

$$r_i = \beta_0 + \sum_{i=1}^k \beta_i K(x_i, \mu_i) \quad (3)$$

where $x_i \in R^d$ and $r_i \in R$ are input training data, $\beta = (\beta_0, \dots, \beta_k) \in R^{k+1}$ is a vector of regression coefficients. K is a local kernel function centered on $\mu \in R^d$. In order to simplify the regression problem, we first perform K-NN clustering algorithm on input training data x_i and assign the kernel function centers μ to be the clustering centers. Suppose r_i , x_i as well as the kernel function K are available, the regression parameter $\beta = (\beta_0, \dots, \beta_k) \in R^{k+1}$ can be solved by a standard least square algorithm according to equation (3).

In our implementation, x_i is the neighbors' low dimensional coordinates Y_{nb} , and r_i is the neighbors' eigen-representations which takes the form:

$$r_i = \begin{pmatrix} U_{nb} \\ y_{nb} \\ \bar{X}_{nb} \end{pmatrix}$$

After β is calculated, given Y_{in} as input, the eigenvectors U_{in} , coefficients y_{in} and mean value \bar{X}_{in} of the expression sequence corresponding to the input image are synthesized. So according to PCA theory, the new expression sequence corresponding to the input neutral face can be reconstructed frame by frame through formula (4):

$$f = U_{in} \cdot y_{in} + \bar{X}_{in} \quad (4)$$

4 Experiments and Discussion

There exists some facial databases for research purpose, such as FERET [14] and AR [15] facial database. The current facial databases may provide face images with variable expression, pose, and illumination, however, much of those available are grayscale images and the facial expression video sequence database is not easy to gain. So we capture facial expression videos by ourselves. We use a Sony HDV 1080i video camera recorder, the video frame resolution amounts to 1920×1080 . To ensure good performance, the actors are informed to perform expressions from neutral to apex with no rigid movements of the head. The distance between human face and the camera is fixed at 2 meters.

Our facial expression database includes 112 video sequences covering 4 kinds of typical expressions (happy, angry, surprise, fear) coming from 28 individuals, each video sequence is normalized to 20 video frames. Since the most sensitive parts of a human face are the eyes and the mouth, we separate the face into upper and lower interest regions and treat them respectively during training and synthesizing, the method for hallucinating the eyes and the mouth is totally identical. We manually cut the upper region and lower region of each video frame for video sample training, a small number of the upper and lower face regions selected from the whole big training set are shown in Fig.1. After implementation of the proposed approach, the hallucinated eyes and mouth are manually transplanted onto the input neutral face frame by frame without any change. To deal with the 24bit true color video frames, our approach is applied on the R, G, B channels respectively, and the final results are the integration of the three channels. We perform the "cross validation" process (randomly pick one out of training data) 10 times and part of the experimental results are demonstrated in Fig.2 and Fig.3. In our experiments, the hallucination procedure generates 20 video frames for each input image in about 10 seconds on a Pentium IV 2.4GHz PC, the kernel function of RBF regression is multi-quadratic function.

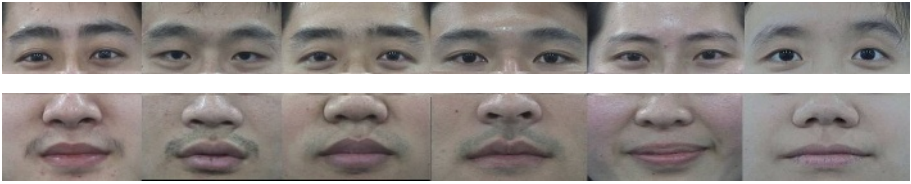


Fig. 1. A number of selected upper and lower regions of different sample faces in our training database. The first line is the upper regions, the second line is the lower regions.

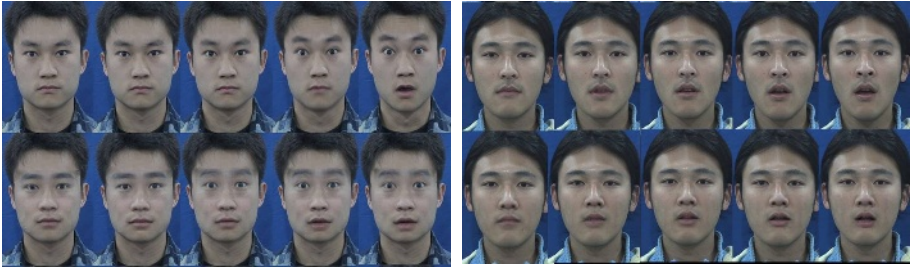


Fig. 2. Surprise expression of two individuals. The first line is the ground-truth facial expression sequences, the second line is hallucinated facial expression sequences.

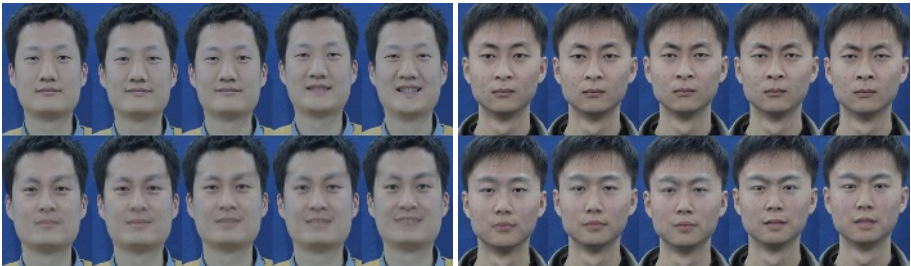


Fig. 3. Happy and angry expressions of two individuals. The first line is the ground-truth facial expression sequences, the second line is hallucinated facial expression sequences. In both line, the left 5 columns are happy expression, the right 5 columns are angry expression.

Compared with the weighted linear combination method adopted in [9], our approach maintains more high frequency information, see Fig.4. In our test, the weights are gained by minimizing equation (1), then the linear combination is performed using the weights and corresponding neighbors frame by frame to synthesize facial expression sequence.

Through the global nonlinear learning, we've computed the low dimensional coordinates of the input subject and the training samples by minimizing equation (2). It's proved that two factors may influence the final hallucination results, i.e., the neighborhood size of LLE and the dimensionality of the low dimensional coordinates. The average RMS (root mean square) error of 10 cross validation tests using different

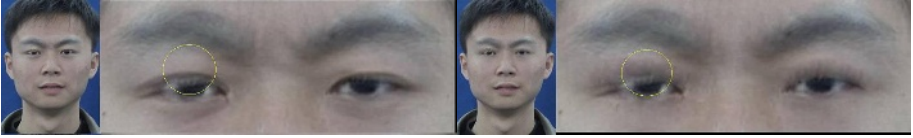


Fig. 4. Hallucinated expression by our approach (left side) and weighted linear combination method (right side). The upper region of the face is magnified on both sides. Note that near the region of the yellow circle, the result by weighted linear combination method looks smoother.

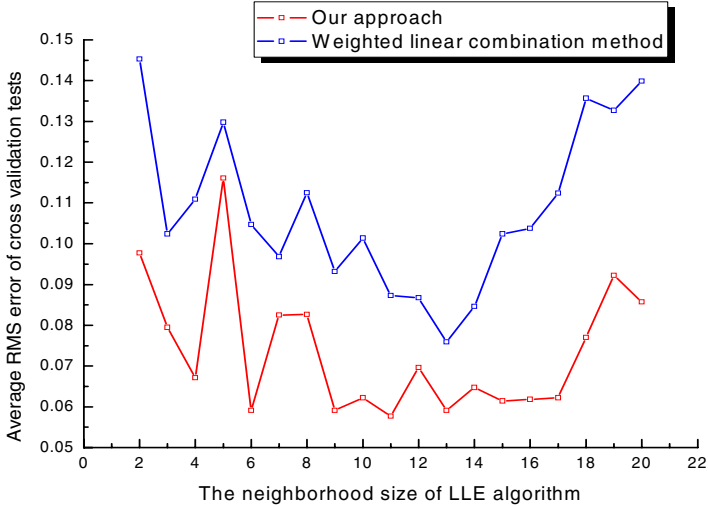


Fig. 5. The average RMS error under different neighborhood size

neighborhood size and dimensionality are shown in Fig.5 and Fig.6. The RMS error is computed according to the only input original image and the first frame of the hallucinated expression sequences. We compare the RMS error of our approach with that of the weighted linear combination method and verify the superiority of our approach.

Fig.5 indicates that the RMS error is very unstable when the neighborhood size is less than 8. We adjust the neighborhood size empirically, when the neighborhood size is between 8 and 17, the RMS error remains relatively stable at lower values, when the neighborhood size surpasses 17, the RMS error rises dramatically.

Fig.6 shows that in global LLE learning, when the dimensionality of the low dimensional coordinates is between 8 and 16, the mean RMS error remains at lower values, otherwise, the RMS error rises dramatically.

Though the neighborhood size and the dimensionality of the low dimensional coordinates do influence the hallucination results, there lacks perfect approach to determine these parameters automatically. In many applications, these parameters are determined empirically according to different cases. In our experiments shown in Fig.2 and Fig.3, the neighborhood size and the dimensionality of the low dimensional coordinates are fixed at 11 and 9 respectively.

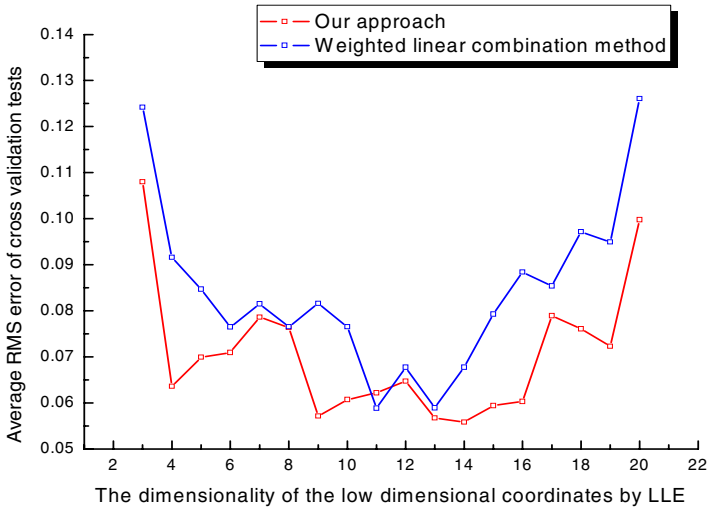


Fig. 6. The average RMS error under different dimensionality of the low dimensional coordinates computed by LLE

5 Conclusion

In this paper, we present a novel two-level hierarchical fusion approach to hallucinate facial expression sequences from training video samples given only one frontal face image with neutral expression. According to the fusion approach, the local linear subspace learning is combined with the global nonlinear subspace learning, the local level simplifies the video hallucination by eigen-representation of the samples in temporal domain, and the global level creates optimized expression appearance in spatial domain. The two-level hierarchical fusion approach provides a sound solution to the problem of organizing the complex training video sample space, and this is the main contribution of our work. Our approach generates reasonable facial expression sequences with little artifact compared with existing method.

References

1. Qingshan, Z., Zicheng, L., Baining, G., Harry, S.: Geometry-Driven Photorealistic Facial Expression Synthesis. Eurographics/SIGGRAPH Symposium on Computer Animation, San Diego, CA (2003) 177-186
2. Zicheng, L., Ying, S., Zhengyou, Z.: Expressive Expression Mapping with Ratio Images. Proceedings of ACM SIGGRAPH. Los Angeles, California (2001) 271-276
3. Simon, B., Takeo, K.: Hallucinating Faces. Proceedings of the Fourth IEEE International Conference on Automatic Face and Gesture Recognition. Grenoble, France (2000) 83-88
4. Ce, L., Heung-Yeung, S., Chang-Shui, Z.: A Two-step Approach to Hallucinating Faces: Global Parametric Model and Local Nonparametric Model. Proceedings of the 2001 IEEE Computer Society Conference on Computer Vision and Pattern Recognition. Kauai Marriott, Hawaii (2001) 192-198

5. Yang, L., Xueyin, L.: An Improved Two-step Approach to Hallucinating Faces. Proceedings of The Third International Conference on Image and Graphics, Hong Kong (2004)
6. Wei, L., Dahua, L., Xiaoou, T.: Face Hallucination Through Dual Associative Learning. IEEE International Conference on Image Processing, Vol. 1, 11-14 (2005) I-873-6.
7. Wei, L., Dahua, L., Xiaoou, T.: Hallucinating Faces: TensorPatch Super-Resolution and Coupled Residue Compensation. Proceedings of IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Vol. 2. (2005) 478-484
8. Wei, L., Dahua, L., Xiaoou, T.: Neighbor Combination and Transformation for Hallucinating Faces. Proceedings IEEE International Conference on Multimedia and Expo.(2005) 145-148
9. Xiaogang W., Xiaoou, T.: Hallucinating Face by Eigentransformation. IEEE Transactions on Systems, Man, and Cybernetics—PART C: Applications and Reviews, vol. 35. (2005) 425-434
10. Congyong, S., Li, H.: Facial Expression Hallucination. Seventh IEEE Workshops on Application of Computer Vision, Vol. 1. (2005) 93-98
11. Matthew, T., Alex, P.: Eigenfaces for Recognition. Journal of Cognitive Neuroscience, 3 , (1991) 71-86
12. Sam, T.R., Lawrence, K.S.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. SCIENCE, 290(5500).(2000) 2323-2326
13. Jaco, V., Simon, J.G., Arnaud, D.: Radial Basis Function Regression Using Trans-dimensional Sequential Monte Carlo. in IEEE Workshop on Statistical Signal Processing (2003)
14. PJonathon, P., Hyeonjoon, M., Patrick, R., Syed, A.R: The FERET Evaluation Methodology for Face Recognition Algorithms. IEEE Transactions on Pattern Analysis and Machine Intelligence, Vol. 22, No. 10 (2000)
15. Martinez, A., Benavente, R.: The AR Face Database. CVC Technical Report #24, (1998)

Blue Sky Detection for Picture Quality Enhancement

Bahman Zafarifar^{2,3} and Peter H.N. de With^{1,2}

¹ Eindhoven University of Technology, PO Box 513, 5600 MB, The Netherlands
{B.Zafarifar, P.H.N.de.With}@tue.nl

² LogicaCMG, PO Box 7089, 5600 JB Eindhoven, The Netherlands

³ Philips Innovative Applications (CE), Pathoekeweg 11, 8000 Burges, Belgium

Abstract. Content analysis of video and still images is attractive for multiple reasons, such as enabling content-based actions and image manipulation. This paper presents a new algorithm and feature model for blue-sky detection. The algorithm classifies the sky areas by computing a pixel-accurate sky probability. Such a probabilistic measure matches well with the requirements of typical video enhancement functions in TVs. This algorithm enables not only content-adaptive picture quality improvement, but also more advanced applications such as content-based annotation of, and retrieval from image and video databases. When compared to existing algorithms, our proposal shows considerable improvements in correct detection/rejection rate of sky areas, and an improved consistency of the segmentation results.

1 Introduction

Sky is among the objects of high visual importance, appearing often in video sequences and photos. A sky-detection system can be used for different applications. At the semantic level, sky detection can contribute to image understanding by e.g. indoor/outdoor classification or automatic detection of image orientation. At this level, applications of sky detection include content-based actions such as image and video selection and retrieval from data-bases, or object-based video coding. At the pixel level, sky detection can be used for content-based image manipulation, like picture quality improvement using color enhancement and noise reduction, or as background detection for 3D depth-map generation.

Content-adaptive processing in general, and sky detection in specific, can be used in high-end televisions. Modern TVs employ a variety of signal-processing algorithms for improving the quality of the received video signal. The settings of these processing blocks are often globally constant or adapted to some local pictorial features, like color or the existence of edges in the direct neighborhood. Such features are often too simple to deal with the diverse contents of video sequences, leading to a sub-optimal picture quality as compared to a system that locally adapts the processing to the content of the image. The above-mentioned local adaptation can be realized if the image is analyzed by a number of object

detectors, after which areas of similar appearance are segmented and processed with algorithms optimized to the features of each area [1].

Due to its smooth appearance, noise and other artifacts are clearly visible in sky regions. This motivates using appropriate image enhancement techniques specifically in the sky regions. The existence of special circuits in high-end TVs for improving the color in the range of sky-blue also illustrates the subjective importance of sky.

Our objective is to develop a sky-detection algorithm, suitable for image enhancement of video sequences. This implies that the detection must be pixel accurate and consistent, and allow for real-time embedded implementation.

Previous work on sky detection includes a system [2][3], based on calculating an initial "sky belief map" using color values¹ and a Neural Network, followed by connected-area extraction. These areas may be accepted or rejected using texture and vertical color analysis, and the degree of fitting to a two-dimensional (2D) spatial model. While this method yields useful results in annotating sky regions, we found it not suitable for the requirements of video applications concerning spatial consistency. The algorithm takes crisp classification decisions per connected-area, leading to abrupt changes in the classification result. As an example, patches of sky may be rejected when their size reduces during a camera zoom-out.

A second system proposed in [4][5] is based on the assumption that sky regions are smooth and are normally found at the top of the image. Using predefined settings, an initial sky probability is calculated based on color, texture and vertical position, after which the settings are adapted to regions with higher initial sky probability. These adapted settings are used for calculating a final sky-probability. The employed pixel-oriented technique (as opposed to the connected-area approach of the first system) makes this system suitable for video applications. However, due to its simple color modeling, this method often leads to false detections, such as accepting non-sky blue objects as sky, and false rejections, like a partial rejection of sky regions when they cover a large range in the color space.

We propose an algorithm that builds upon the above-mentioned second system, and exploits its suitability for video applications, while considerably improving the false detection/ rejection rates. The proposed sky detector is confined to blue-sky regions, which includes both clear blue sky and blue sky containing clouds.

Experimental simulations of our new proposal indicate a substantial improvement in the correct detection of sky regions covering a large color range, and the correct rejection of non-sky objects when compared to the algorithm proposed in [4][5], as well as an improved spatial consistency with respect to the system described in [2][3].

¹ In this paper, "color" denotes all color components. When a distinction between chromaticity and gray values is required, we use the terms "luminance" and "chrominance".



Fig. 1. Various appearances of sky, from left to right: dark, light, large color range, occluded

The remainder of the paper is organized as follows. Section 2 characterizes the sky features, Section 3 describes the proposed algorithm, Section 4 presents the results and Section 5 concludes the paper.

2 Observation of Sky Properties

In this section, we discuss the features of sky images, and address the challenges for modeling the sky.

Sky can have a variety of appearances, such as clear sky, cloudy sky, and overcast sky (see Fig. 1). Sky color can cover a large part of the color space, from saturated blue to gray, or even orange and red during sun-set. Consequently, a system based on temporally-fixed color settings is likely to fail in correctly detecting different sky appearances. In addition, sky regions can significantly vary in color within an image: a wide-shot clear-sky image tends to be more saturated at the top and becomes less saturated near the horizon, while the luminance tends to increase from the top of the image towards the horizon. As a result, a sky detector using a spatially-fixed color is likely to reject parts of the sky region, when the sky color considerably changes within one image.

An additional challenge is the partial occlusion of sky by foreground objects, cutting the sky into many disconnected parts. In order to prevent artifacts in the post-processed video, it is important that all sky areas are assigned coherent probabilities.

Another non-trivial task is distinguishing between sky, and objects which look similar to sky but are actually not a part of it. Examples are areas of water, reflections of sky, or other objects with similar color and texture as sky.

In the following section, we propose a system that addresses the aforementioned issues.

3 Algorithm Description

3.1 Sky Detector Overview

We propose a sky-detection system based on the observation that blue-sky regions are more likely to be found at the top of the image, they cover a certain part of the color space, have a smooth texture, and the pixel values show limited horizontal and vertical gradients. The algorithm contains three stages, as depicted in Fig. 2.

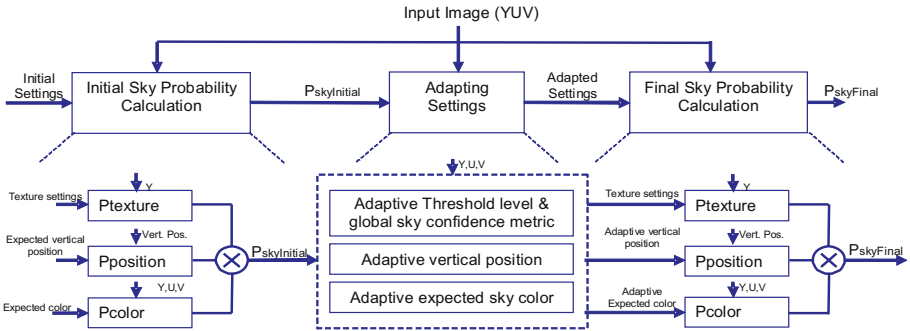


Fig. 2. Block diagram of the Sky detector, divided in three stages

Stage 1 : Initial sky probability. In this stage, an initial sky-probability map is calculated based on the color, vertical position and texture of the image pixels. The texture analysis also includes horizontal and vertical gradient measures. The settings of this stage are fixed, and are chosen such that all targeted sky appearances can be captured.

Stage 2 : Analysis and sky-model creation. In this stage, the fixed settings of the first stage are adapted to the image under process. As such, the settings for the vertical-position probability and the expected sky color are adapted to the areas with high sky probability. For the expected color, a spatially-varying 2D model is created that prescribes the sky color for each image position.

Stage 3 : Final sky probability. In this stage, a pixel-accurate sky probability is calculated based on the color, vertical position and (optionally) texture of the image, using the adaptive model created in Stage 2.

With respect to implementation, we have adopted the YUV color-space because the sky chrominance components in the vertical direction of the image, tend to traverse linearly in the UV plane from saturated blue through gray to red. In order to reduce the amount of computations, the image is down-scaled to QCIF resolution for usage in Stage 1 and 2. However, Stage 3 uses the image at the original resolution in order to produce pixel-accurate results.

Sections 3.2, 3.3 and 3.4 describe the three stages of the algorithm in more detail.

3.2 Initial Sky Probability

Using predefined settings, an initial sky probability ($P_{skyInitial}$) is calculated using a down-scaled version of the image. We combine color, vertical position, and texture to compute the initial sky probability as

$$P_{skyInitial} = P_{color} \times P_{position} \times P_{texture} .$$

1 . The color probability is calculated using a three-dimensional Gaussian function for the Y, U and V components, centered at predetermined positions Y_0 , U_0 and V_0 (representing the expected sky color), with corresponding standard deviations σ_{y1} , σ_{u1} and σ_{v1} . The settings are chosen such that all desired sky appearances are captured. The color probability is defined as

$$P_{color} = e^{-\left(\left(\frac{Y-Y_0}{\sigma_{y1}}\right)^2 + \left(\frac{U-U_0}{\sigma_{u1}}\right)^2 + \left(\frac{V-V_0}{\sigma_{v1}}\right)^2\right)} .$$

2 . The vertical-position probability is defined by a Gaussian function, which has its center at the top of the image, starting with unity value and decreasing to 0.36 at the bottom of the image, so that

$$P_{position} = e^{-\left(\frac{r}{height}\right)^2} ,$$

where r is the vertical coordinate of the current pixel (at the top of the image $r = 0$) and $height$ denotes the total number of rows (i.e. TV lines) of the image.

3 . The calculation of the texture probability is based on a multi-resolution analysis of the luminance channel of the image. The analysis assigns low probabilities to parts of the image containing high luminance variation, or excessive horizontal or vertical gradients. This probability can be used to eliminate the textured areas from the initial sky probability. More specifically, three down-scaled (with factors of 2) versions of the luminance channel are analyzed using a fixed window-size (of 5×5 pixels), and the results are combined in the lowest resolution, using the *minimum* operator. The texture analysis uses the following two measures.

SAD: The local smoothness of the image can be measured by the luminance variation. Using the Sum of Absolute Differences (SAD) between horizontally-adjacent, and vertically-adjacent pixels in the analysis window, we calculate the luminance variation in the surrounding of the current pixel. The horizontal and vertical *SAD* (SAD_{hor} and SAD_{ver}) lead to a probabilistic measure P_{SAD} as follows

$$SAD_{hor}(r, c) = \frac{1}{N_{SAD}} \sum_{i=-w}^w \sum_{j=-w}^{w-1} |Y(r+i, c+j) - Y(r+i, c+j+1)| ,$$

$$SAD_{ver}(r, c) = \frac{1}{N_{SAD}} \sum_{i=-w}^{w-1} \sum_{j=-w}^w |Y(r+i, c+j) - Y(r+i+1, c+j)| ,$$

$$P_{SAD} = e^{-\left([SAD_{hor} + SAD_{ver} - T_{SAD}]_0^\infty\right)^2} .$$

Here, r and c are the coordinates of the pixel in the image, w defines the size of the analysis window (window size = $2w + 1$), and i and j are indices of the window. The factor $1/N_{SAD}$ is used to normalize the SAD to the total number of the pixel differences within the window ($N_{SAD} = (2w + 1) * 2w$), and T_{SAD} is a noise-dependent threshold level. The symbol $[\cdot]_0^\infty$ denotes a clipping function defined as

$$[f]_a^b = \text{Min}(\text{Max}(f, a), b).$$

Gradient: we observe that luminance values of the sky regions have limited horizontal and vertical gradients, and that the luminance often increases in top-down direction. We define the vertical gradient ($grad_{ver}$) as the difference between the sum of pixel values of the upper-half of the analysis window, and the sum of the pixel values of the lower-half of the analysis window. The horizontal gradient ($grad_{hor}$) is defined similarly, using the pixels of the left-half and the pixels on the right-half of the analysis window. For pixel coordinate (r, c) this leads to

$$grad_{hor}(r, c) = \frac{1}{N_{grad}} \left(\sum_{i=-w}^w \sum_{j=-w}^{-1} Y(r+i, c+j) - \sum_{i=-w}^w \sum_{j=1}^w Y(r+i, c+j) \right),$$

$$grad_{ver}(r, c) = \frac{1}{N_{grad}} \left(\sum_{i=-w}^{-1} \sum_{j=-w}^w Y(r+i, c+j) - \sum_{i=1}^w \sum_{j=-w}^w Y(r+i, c+j) \right),$$

where the factor $1/N_{grad}$ normalizes the gradient to the size of the window ($N_{grad} = w * (2w + 1)$).

Using appropriate threshold levels, the horizontal and vertical gradients are translated to a probability P_{grad} , calculated as

$$P_{grad} = e^{-([\text{ } T_{vl} - grad_{ver}]_0^\infty + [grad_{ver} - T_{vu}]_0^\infty + [|grad_{hor}| - T_h]_0^\infty)^2},$$

where T_{vl} and T_{vu} are the threshold levels for the lower and upper bounds of the vertical gradient respectively, and T_h is the threshold level for the horizontal gradient. These thresholds are fixed values, determined by a set of training images. Using separate thresholds for the upper and lower bounds in the vertical direction allows an increase, and penalized a decrease of the luminance in the downwards image direction.

Finally, the texture probability $P_{texture}$ combines P_{SAD} and P_{grad} as

$$P_{texture} = P_{SAD} \times P_{grad} .$$

3.3 Analysis and Sky-Model Creation

In this stage, the initial sky probability (calculated in Stage 1) is analyzed in order to create adaptive models for the color and vertical position used in the final sky-probability calculation. This involves the following steps.

1. *Calculating Adaptive threshold level and global sky confidence metric:* the initial sky probability needs to be segmented in order to create a map of regions with high probability. Simple measures for threshold determination, such as using the maximum of the sky-probability map as proposed in [5] can perform inadequately, for example by favoring small objects with high sky probability over larger non-perfect sky regions. In order to avoid this problem, we propose a more robust method that takes both the size and the probability of sky regions into account, by computing an *adaptive* threshold and a global sky confidence-metric Q_{sky} . The confidence metric yields a high value if the image contains a significant number of pixels with high initial sky probability. This prevents small sky-blue objects from being accepted as sky, in images where no large areas with high sky probability are present. The calculation steps are as follows: first the Cumulative Distribution Function (CDF) of the initial sky probability is computed, after which it is weighted using a function that emphasizes the higher sky probability values and decreases to zero towards the lower sky probability values. Due to this weighting, the position of the maximum of the resulting function (weighted CDF) includes our preference for higher probability values, while being dependent on the distribution of the initial sky probability values. Therefore, this position can be used to determine the desired adaptive threshold. The maximum amplitude of the weighted CDF is dependent on the number of pixels with relatively high sky probability, and thus can be used for determining the aforementioned confidence metric Q_{sky} .

2. *Adaptive vertical position:* the areas with high sky-probability are segmented by thresholding the initial sky-probability map, with the threshold level described in the previous paragraph, after which the mean vertical position of the segmented areas is computed. This adaptive vertical position is used to define a function, which equals unity at the top of the image and linearly decreases towards the bottom of segmented sky region. This function is then used for computing the final sky probability.

3. *Adaptive expected sky color:* as mentioned in Section 2, the sky detector needs to deal with the wide range of sky color values within and between different frames. In [5], it is proposed to use frame-adaptive, but further spatially-constant expected colors. This method addresses the problem of large color variation between frames, but fails when the sky covers a considerable color range within one frame, resulting in a partial rejection of the sky areas.

To address this problem, we propose to use a spatially-adaptive expected sky color. To this end, each signal component (Y, U, and V) is modeled by a spatially-varying 2D function, that is fitted to a selected set of pixels with high sky probability.

An example of model fitting technique is as follows. Using a proper adaptive threshold, the initial sky probability is segmented to select sky regions with high sky probability. Next, the segmented pixels are selected with a decreasing density in top-down direction. This exploits our assumption that the pixels at the top are more important for model fitting than those near the bottom, and ensures that the model parameters are less influenced by reflections of sky or

other non-sky blue objects below the actual sky region. The last step is to use the values of the (Y, U, V) signal components of these selected pixels to fit the 2D function of the corresponding signal component.

The choice of the color model and the fitting strategy of the 2D functions depend on the required accuracy and the permitted computational complexity. We implemented (1) a 2D second-degree polynomial, in combination with a least-squares optimization for estimating the model parameters, and (2) a model which uses a matrix of 23×18 values, per color component, for representing the image color [6]. The 2^{nd} -degree polynomial model offers sufficient spatial flexibility to represent typical sky colors, but is computationally expensive (the presented results in this paper use this model). The second model, also offers the necessary flexibility, and is in addition more suitable for hardware implementation.

3.4 Final Sky Probability

Using the adaptive model created in Stage 2, we compute a pixel-accurate final sky-probability map as

$$P_{skyFinal} = P_{color2} \times P_{position2} \times P_{texture2} \times Q_{sky} ,$$

where Q_{sky} denotes the sky confidence metric. The required pixel accuracy is achieved by using the original image resolution, and applying a moderate texture measure to prevent distortion in the final sky probability map, near the edges of non-sky objects. The following paragraphs further describe the features applied in this stage.

1. The color probability is calculated using a 3D Gaussian function for Y, U and V components, centered at the spatially-varying values $Y_{0,(r,c)}$, $U_{0,(r,c)}$ and $V_{0,(r,c)}$ (representing expected sky color at the spatial position (r, c)), with corresponding standard deviations σ_{y2} , σ_{u2} and σ_{v2} . In order to reduce false detections, these standard deviations are reduced with respect to the values of stage 1.

2. As opposed to the fixed vertical-position function used for initial sky probability, the final stage uses an adaptive vertical probability function, which is tuned to cover the sky areas with high sky probability, as calculated in Stage 2.

3. The inclusion and the type of texture measure depend on the application for which the sky detection output is used. For some applications, using a texture measure in the final sky-probability calculation could lead to undesirable effects in the post-processed image, while other applications may require some form of a texture measure. For example, for noise removal in the sky regions, we found it necessary to reduce the sky probability of pixels around the edges of objects adjacent to sky, in order to retain the edge sharpness. This was done by taking the Sobel edge detector as texture measure.

4 Experimental Results

We applied the proposed algorithm on more than 200 sky images. The images were selected to present a large variety of sky appearances, many including sky



Fig. 3. Examples of improved correct detection, left: input, middle: proposed by [4][5], right: our algorithm



Fig. 4. Examples of improved correct rejection, left: input, middle: proposed by [4][5], right: our algorithm

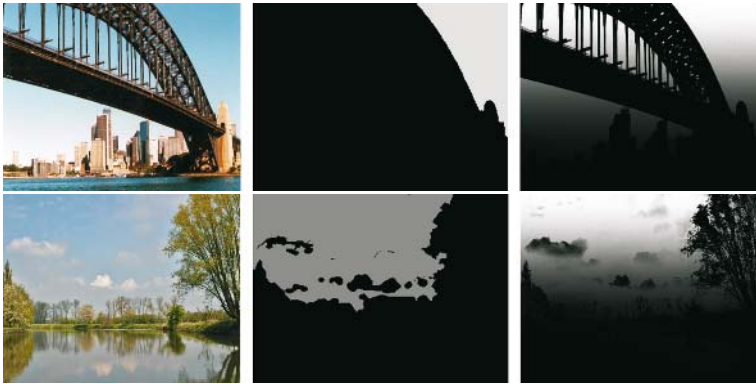


Fig. 5. Examples of improved spatial accuracy, left: input, middle: proposed by [2][3] (courtesy of Eastman Kodak), right: our algorithm

reflections and other challenging situations. Figure 3 compares our results to [4][5]. In Fig. 3-top, the halo (top-middle) is resulted by the spatially constant color model used in [4][5], while the spatially adaptive color model employed in our algorithm is capable of dealing with the large color range of the sky area (top-right). A similar difference in the results can be seen in Fig. 3-bottom, where in addition, the reflection of the sky is removed by the gradient analysis. Figure 4 shows the improved correct rejection of non-sky objects (areas or water in Fig. 4-top and mountains in Fig. 4-bottom), which has been achieved because of the multi-resolution texture analysis. Lastly, Fig. 5 shows the greatly improved spatial accuracy of our results in comparison to [2][3]. This is due to the two-pass approach for calculating the sky probability, in which the second pass uses the original image resolution and a moderate texture measure.

When compared to [4][5], our experiments indicate a substantial improvement in the correct detection of sky regions covering a large color range due to the spatially varying color model, and an improved correct rejection of non-sky objects due to the multi-resolution texture analysis.

When compared to [2][3], we observed an improved spatial consistency of the segmentation results. Here, a notable improvement in the correct detection was discovered in 16 out of 23 images, where a side-by-side visual comparison was made. In the remaining cases, our proposal performed comparable to the existing system. In many of these cases we still prefer our proposal, as it is based on a smooth probability measure, whereas the existing system produces crisp results, which is more critical in the case of false detections for video applications. An experiment with a video sequence indicated that the spatial consistency also improves the temporal behavior of the system. More algorithmic tuning and experiments will have to be conducted to validate this conjecture.

A simplified version of the proposed algorithm is currently being implemented as a real-time embedded system, using FPGA technology. Preliminary mapping results indicate that a real-time implementation is feasible on a standard FPGA device.

5 Conclusions

Sky detection for video sequences and still images can be used for various purposes, such as automatic image manipulation (e.g. picture quality improvement) and content-based directives (e.g. interactive selection and retrieval from multimedia databases). The main problems with the existing algorithms is incomplete detection of sky areas with large ranges of color, false detection of sky reflections or other blue objects, and inconsistent detection of small sky areas. This paper has presented a sky-detection algorithm which significantly reduces the mentioned problems, and has suitable properties for video applications. This was achieved by constructing a sky model that incorporates a 2D spatially-varying color model, while reusing the vertical position probability from an existing method. Moreover, we have introduced a confidence metric for improving the consistency and removal of small blue objects. Wrong detection of the reflections

of sky areas and other non-sky objects has been reduced by employing a gradient analysis of the luminance component of the sky. Experimental results show that the proposed algorithm is capable of handling a broad range of sky appearances. The two primary advantages of the proposed algorithm are increased correct detection/rejection rates, and an improved spatial accuracy and consistency of the detection results.

Our future work includes developing additional measures for meeting the requirements of real-time video applications. Particularly, the key parameters of the system, such as the vertical position model, the color model, and the confidence metric need to be kept consistent over time. Furthermore, the algorithm will be optimized for implementation in consumer television systems.

Acknowledgement

The authors gratefully acknowledge Dr. Erwin Bellers for his specific input on existing algorithms. We are also thankful to Dr. Jiebo Luo for providing us with the results of sky-detection algorithm described in [2][3] on a number of sample images.

References

1. S. Herman and J. Janssen, “*System and method for performing segmentation-based enhancements of a video image*”, European Patent EP 1 374 563, date of publication: January 2004.
2. A.C. Gallagher, J. Luo, and W. Hao, “Improved blue sky detection using polynomial model fit,” in *IEEE International Conference on Image Processing*, October 2004, pp. 2367–2370.
3. J. Luo and S. Etz, “*Method for detecting sky in images*”, European Patent EP 1 107 179, date of publication: February 2001.
4. S. Herman and E. Bellers, “Locally-adaptive processing of television images based on real-time image segmentation,” in *IEEE International Conference on Consumer Electronics*, June 2002, pp. 66–67.
5. S. Herman and E. Bellers, “*Adaptive segmentation of television images*”, European Patent EP 1 573 673, date of publication: September 2005.
6. P.H.N. de With B. Zafarifar, “Adaptive modeling of sky for video processing and coding applications,” in *27th Symposium on Information Theory in the Benelux*, June 2006, pp. 31–38.

Requantization Transcoding in Pixel and Frequency Domain for Intra 16x16 in H.264/AVC

Jan De Cock, Stijn Notebaert, Peter Lambert,
Davy De Schrijver, and Rik Van de Walle

Department of Electronics and Information Systems – Multimedia Lab
Ghent University – IBBT

Gaston Crommenlaan 8 bus 201, B-9050 Ledeborg-Ghent, Belgium
`jan.decock@ugent.be`

Abstract. In the context of Universal Multimedia Access, efficient techniques are needed for the adaptation of video content. An important example is the reduction of the bitrate in order to satisfy the bandwidth constraints imposed by the network or the decoding capability of the terminal devices. Requantization transcoding is a fast technique for bitrate reduction, and has been successfully applied in previous video coding standards such as MPEG-2. In this paper, we examine requantization in H.264/AVC, focusing on the intra 16×16 prediction modes. Due to the newly introduced coding tools in H.264/AVC, new techniques are needed that are able to lower the bitrate at a minimal quality loss. We propose two novel architectures, one in the pixel domain and one in the frequency domain, that reuse the information from the incoming bitstream in an efficient way, and perform approximately equally well as a cascade of decoder and encoder. Due to their low computational complexity, the introduced architectures are highly suitable for on-the-fly video adaptation scenarios.

1 Introduction

More and more video for multimedia applications is coded using the H.264/AVC standard [1]. For distribution of this video content, the network and the terminals impose varying constraints on the characteristics of the transferred video bitstreams. In this context, the bitstreams have to be adapted, and a reduction of the framerate, the spatial resolution or the bitrate is required. The latter is possible by means of requantization of the initial bitstreams. In order to perform this requantization, different architectures are possible. The most straightforward solution is the cascade of a decoder-encoder pair. Due to its high computational complexity, this cascade is less eligible for real-time adaptation scenarios [2,3]. Hence, new architectures are needed, that are able to adapt H.264/AVC bitstreams in an efficient way. Transcoding is a fast and elegant solution for bitrate reduction, making it possible to change the characteristics of video sequences without fully decoding and re-encoding [2,3]. Requantization

transcoding has been applied for previous coding standards, such as MPEG-1 and MPEG-2 Video [4,5].

In this paper, we examine requantization transcoding for H.264/AVC, and revise the existing open-loop architecture. Due to the newly introduced coding tools, such as H.264/AVC intra prediction, an extension of the existing architectures is required. Because of the improved coding efficiency of H.264/AVC and the high amount of dependencies in the video bitstreams, the transcoded video sequences are highly susceptible to quality degradation due to drift propagation. In order to constrain the quality loss, two novel closed-loop transcoding architectures for requantization of intra coded pictures are presented. These architectures focus on the H.264/AVC intra 16×16 prediction, and compensate for the prediction errors originated by the requantized prediction pixels. One architecture performs this compensation in the pixel domain, the other in the transform domain.

The remainder of this paper is organized as follows. In Sect. 2, an overview of the 16×16 intra prediction, the transformations, and the quantization of the H.264/AVC standard is given. In Sect. 3, an architecture for an open-loop requantization transcoder targeting H.264/AVC bitstreams is described. Sections 4 and 5 present two novel drift-reducing architectures for requantization transcoders exploiting pixel-domain and frequency-domain compensation methods. In Sect. 6, these types of transcoders are compared based on quality and bitrate measurements of transcoded H.264/AVC bitstreams. Finally, Sect. 7 concludes this paper.

2 H.264/AVC Tools

2.1 Intra Prediction

Intra prediction is used to exploit the spatial redundancy between neighbouring pixels. A block is predicted using previously encoded and reconstructed pixels of surrounding blocks. In H.264/AVC, a macroblock can be predicted using a combination of nine 4×4 or one of four 16×16 intra prediction modes. The intra prediction, which was not present in, for example, MPEG-2 Video, results in an improved compression efficiency. However, it also introduces a number of dependencies. As we will see, this has an important impact on the perceptual quality of the transcoded video sequences.

In this paper, we focus on the intra 16×16 prediction. The four intra 16×16 modes (vertical, horizontal, DC, and plane prediction) are shown in Fig. 1.

2.2 H.264/AVC Transform and Quantization

The integer transform in the H.264/AVC specification [6,7,8] is based on the Discrete Cosine Transform, and is applied on 4×4 blocks. The forward transform of a 4×4 block X is represented by

$$Y = (C_F X C_F^T) \otimes E_F ,$$

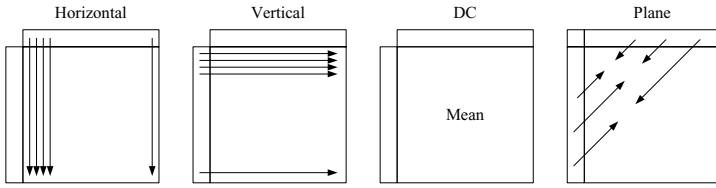


Fig. 1. Intra 16x16 prediction modes

where C_F represents the kernel transformation matrix. E_F is the post-scaling matrix. For efficiency reasons, the post-scaling operation of the transformation is postponed and integrated in the quantization process. First, this forward transform is applied on every 4×4 block of the macroblock. Of these 16 transformed blocks, the DC values are collected in a new 4×4 block, D . On this block, a Hadamard transform is applied:

$$U = (HDH^T)/2.$$

After the core transformation $W_{ij} = (C_F X C_F^T)_{ij}$, with $i, j = 0, \dots, 3$, and the Hadamard transform, the coefficients W_{ij} and U_{ij} are quantized. H.264/AVC provides 52 values for the Quantization Parameter (QP), which can vary on a macroblock basis. The values of QP were defined in such a way that, if QP is increased with a value of 6, the quantization step is doubled and the bitrate is approximately halved. This non-linear behaviour results in the possibility to target a broad range of bitrates. The forward quantization can be implemented as

$$|Z_{ij}| = (|W_{ij}| \cdot M_{ij} + f) \gg \text{qbits}$$

where $\text{qbits} = 15 + \lfloor QP/6 \rfloor$, and f represents the dead zone control parameter [7]. The multiplication factor M_{ij} is determined by $QP \bmod 6$ and the position in the 4×4 block. The quantization for the Hadamard coefficients U_{ij} is performed in a similar way, resulting in the coefficients S_{ij} .

At the decoder side, the process is defined as follows. Before the inverse quantization the inverse Hadamard is applied:

$$U' = H^T S H.$$

For the AC coefficients, the inverse quantization process is defined as

$$W'_{ij} = Z_{ij} \cdot V_{ij} \cdot 2^{\lfloor QP/6 \rfloor}.$$

The values of M_{ij} and V_{ij} result in the coefficients W'_{ij} and D'_{ij} that exceed the pre-quantized values W_{ij} and D_{ij} by a factor $64 \cdot E_{F_{ij}} \cdot E_{I_{ij}}$, hence including the post-scaling of the forward transform along with the pre-scaling of the inverse transform:

$$X' = C_I^T (Y \otimes E_I) C_I.$$

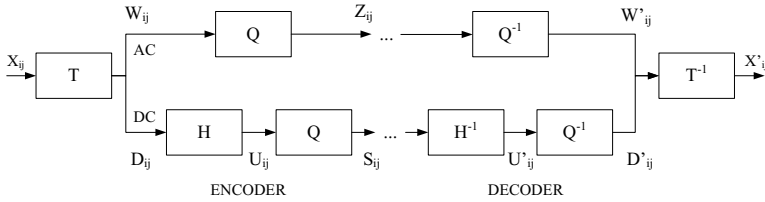


Fig. 2. Transformation and quantization for intra 16×16 macroblocks

The factor 64 is introduced to avoid rounding errors in the inverse transformation that follows. After inverse quantization, the resulting DC coefficients D'_{ij} are placed back in the $16 \ 4 \times 4$ blocks, and the inverse transform is applied.

The entire process of transformation and quantization for intra 16×16 prediction is shown in Fig. 2. We refer to [6,7,8] for more information about the intertwined transformation and quantization.

3 Open-Loop Requantization Transcoder

The most straightforward architecture for requantization is the open-loop transcoder. It consists of a dequantization (Q_1^{-1}), followed by a requantization (Q'_2 and Q''_2) with a coarser QP. The architecture of the open-loop transcoder is visualized in Fig. 3.

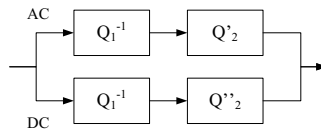


Fig. 3. The open-loop requantization transcoder

The implementation of this type of requantization transcoder was rather straightforward in MPEG-2 [5]. However, in H.264/AVC, special attention has to be paid to the requantization Q'_2 and Q''_2 . The multiplication factors have to be adapted in order to bring into account the scaling factors $E_{F_{ij}}$ and $E_{I_{ij}}$ of the H.264/AVC integer transform. Since these scaling factors are already applied in the original quantization, they may not be repeated in the requantization. Because two types of transformations (the integer and the Hadamard transformation) are used for intra 16×16 prediction, we have to consider both transformations separately in order to perform a correct adaptation of the different 4×4 blocks. As a result, the multiplication factors for the integer transformation

have to be downscaled by the factors 4, 2.56 and 3.2, depending on their position (i, j) in the 4×4 block of coefficients, as shown in Table 1 and Table 2. The downscaling factors arise from:

$$(M_{ij} \cdot V_{ij}) \gg 15 = 64 \cdot E_{F_{ij}} \cdot E_{I_{ij}} = \begin{cases} 4, & r = 0 \\ 2.56, & r = 1 \\ 3.2, & r = 2 \end{cases}$$

where the factor 64 is introduced to avoid rounding errors during the inverse transform¹, and

$$r = \begin{cases} 0, & (i, j) \in \{(0, 0), (0, 2), (2, 0), (2, 2)\} \\ 1, & (i, j) \in \{(1, 1), (1, 3), (3, 1), (3, 3)\} \\ 2, & \text{otherwise} \end{cases}$$

The multiplication factors for the Hadamard transformation have to be scaled as well. In this case, the obtained scaling factor is position-independent. Deriving the factor from the forward and inverse quantization formulas for Hadamard coefficients, we obtain:

$$(M_{0,0} \gg 16) \cdot (V_{0,0} \gg 2) = \frac{1}{2}.$$

This means that the transcoded coefficients have to be upscaled by a factor two, in order to obtain a correct requantization Q''_2 of the DC coefficients.

Table 1. Original multiplication factors M_{ij} and V_{ij} for the integer transform

Table 2. Modified multiplication factor M'_{ij} for the integer transform

QP mod 6	M_{ij}			V_{ij}		
	$r = 0$	$r = 1$	$r = 2$	$r = 0$	$r = 1$	$r = 2$
0	13107	5243	8066	10	16	13
1	11916	4660	7490	11	18	14
2	10082	4194	4194	13	20	16
3	9362	3647	3647	14	23	18
4	8192	3355	3355	16	25	20
5	7282	2893	2893	18	29	23

QP mod 6	M'_{ij}		
	$r = 0$	$r = 1$	$r = 2$
0	3277	2048	2521
1	2979	1820	2341
2	2521	1638	2048
3	2341	1425	1820
4	2048	1311	1638
5	1821	1130	1425

4 Requantization Transcoder with Pixel-Domain Compensation (PDC)

The open-loop transcoder of the previous section is a fast technique for rate reduction. However, because there is no feedback loop, the quality of the outgoing

¹ After the inverse transform, the residual values are downscaled by 64.

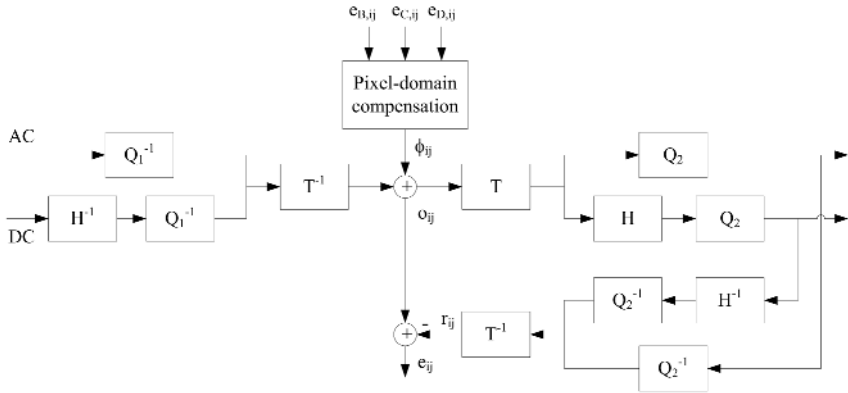


Fig. 4. Transcoder with pixel-domain compensation

bitstream will degrade because of drift propagation. In this section, a novel architecture is presented which reduces the negative impact of requantization. This architecture avoids error propagation by compensating using a mode-dependent matrix, ϕ . The compensation matrix models the effect of requantization differences of surrounding prediction pixels on the prediction of the current 16×16 block. The transcoder architecture is depicted in Fig. 4.

ϕ is constructed as follows. We define the error values e_{ij} as the difference between the incoming residual information after inverse quantization, inverse transformation and drift compensation, and the corresponding requantized residual values after inverse quantization and inverse transformation, i.e., $e_{ij} = o_{ij} - r_{ij}$, for $i, j = 0, \dots, 15$. We assign the error values $e_{B,i,j}$, $e_{C,i,j}$, $e_{D,i,j}$ as e_{ij} for macroblocks B (top), C (left), and D (upper-left), for the current macroblock A. For clarity, we only mention the error values that are required for the construction of the compensation, namely the error values $e_{B,15,j}$ for $j = 0..15$, $e_{C,i,15}$ for $i = 0..15$, and $e_{D,15,15}$. From these 33 values, the pixel-domain compensation matrix ϕ is constructed using the formulas for the intra 16×16 prediction, just as they are used in the encoder and decoder, but here applied on the smaller error values. For example, for horizontal prediction (mode 1) for the four 4×4 blocks in the top row of the macroblock, ϕ becomes:

$$\phi = \begin{bmatrix} e_{C,0,15} & e_{C,0,15} & e_{C,0,15} & e_{C,0,15} \\ e_{C,1,15} & e_{C,1,15} & e_{C,1,15} & e_{C,1,15} \\ e_{C,2,15} & e_{C,2,15} & e_{C,2,15} & e_{C,2,15} \\ e_{C,3,15} & e_{C,3,15} & e_{C,3,15} & e_{C,3,15} \end{bmatrix}.$$

For the DC prediction, the average of the available surrounding pixels has to be calculated once. The plane prediction, however, is more complex and requires two multiplications for every position in the 16×16 macroblock [1].

The advantages of the PDC architecture when compared to a cascaded decoder-encoder pair is that no exhaustive search for the prediction mode is required, since

the mode is passed on from the incoming bitstream. Additionally, the intra prediction formulas have to be applied only once, since we are working on differences of residual values. In a decoder-encoder architecture, the formulas have to be applied once at the decoder side, on the original values (resulting in a prediction matrix P), and at least once² at the encoder side, on the requantized values (resulting in a prediction matrix P'). Because of the linear construction of the prediction formulas and the H.264/AVC integer and Hadamard transformations, the formulas can be applied on the error values e_{ij} directly, resulting in the matrix $\phi = P - P'$. It should be noted, however, that the downscaling after the inverse integer transformation (as mentioned in Sect. 3) and the divisions in the DC and plane prediction modes could result in rounding errors.

5 Requantization Transcoder with Transform-Domain Compensation (TDC)

The pixel-domain transcoder as described in the previous section tries to overcome the quality-related problems of open-loop requantization. The question remains if it possible to reduce the computational complexity of the closed-loop architecture. This reduction is possible by eliminating the forward and inverse transforms, hence working as much as possible in the transform domain. The compensation technique as used in the previous section can also be used in the transform domain. It is possible to calculate the formulas for intra prediction directly in the transform domain, by combining the pixel domain intra prediction formulas and the forward integer and Hadamard transforms. As in the pixel domain, the DC and the AC coefficients are treated separately. In order to calculate the prediction errors e_{ij} , one inverse Hadamard for every 16×16 macroblock and one inverse integer transform for every 4×4 block is still required. This is illustrated in Fig. 5.

For all four intra 16×16 prediction modes, the compensation matrices can be calculated, both for the DC and AC values. This results in a compensation matrix Φ for every 4×4 block, and one matrix Θ to compensate the matrix of Hadamard-transformed DC coefficients. For the example of horizontal prediction in the previous section, the compensation matrix Φ for the four 4×4 blocks in the top row of the macroblock becomes:

$$\Phi = 4 \begin{bmatrix} 0 & 0 & 0 & 0 \\ 2e_{C,0,15} + e_{C,1,15} - e_{C,2,15} - 2e_{C,3,15} & 0 & 0 & 0 \\ e_{C,0,15} - e_{C,1,15} - e_{C,2,15} + e_{C,3,15} & 0 & 0 & 0 \\ e_{C,0,15} - 2e_{C,1,15} + 2e_{C,2,15} - e_{C,3,15} & 0 & 0 & 0 \end{bmatrix}.$$

and Θ :

$$\Theta = 8 \begin{bmatrix} \alpha + \beta + \gamma + \delta & 0 & 0 & 0 \\ \alpha + \beta - \gamma - \delta & 0 & 0 & 0 \\ \alpha - \beta - \gamma + \delta & 0 & 0 & 0 \\ \alpha - \beta + \gamma - \delta & 0 & 0 & 0 \end{bmatrix}.$$

² Multiple times, in case of an exhaustive search for the optimal prediction mode.

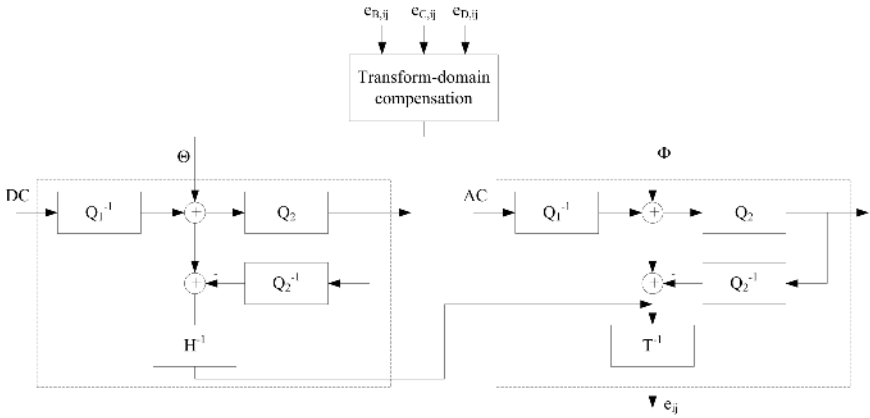


Fig. 5. Transcoder with transform-domain compensation

where

$$\alpha = \sum_{i=0}^3 e_{C,i,15}, \quad \beta = \sum_{i=4}^7 e_{C,i,15}, \quad \gamma = \sum_{i=8}^{11} e_{C,i,15}, \quad \delta = \sum_{i=12}^{15} e_{C,i,15} .$$

For the DC prediction mode, it suffices to compensate the matrix of DC values in only one position, and no compensation is needed for the AC coefficients. In this case, Θ becomes:

$$\Theta = 4 \begin{bmatrix} \sum_{i=0}^{15} (e_{B,15,i} + e_{C,i,15}) & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix} .$$

Computationally, the most complex mode is the plane prediction mode. If we transform the H.264/AVC prediction formulas, we obtain the following. As in the pixel domain, let a , b , and c be defined as:

$$\begin{aligned} a &= 64(e_{B,15,15} + e_{C,15,15}) \\ b &= 5 \left[8(e_{B,15,15} - e_{D,15,15}) + \sum_{i=0}^6 (i+1)(e_{B,15,6-i} - e_{B,15,8+i}) \right] \\ c &= 5 \left[8(e_{C,15,15} - e_{D,15,15}) + \sum_{j=0}^6 (j+1)(e_{C,6-j,15} - e_{C,8+j,15}) \right] . \end{aligned}$$

Further, if we define

$$\begin{aligned} b_2 &= b/512 \\ c_2 &= c/512 , \end{aligned}$$

we obtain the transform-domain Hadamard compensation matrix as follows:

$$\Theta = \begin{bmatrix} a - 3/32b - 3/32c & b/4 & 0 & b/8 \\ c/4 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ c/8 & 0 & 0 & 0 \end{bmatrix}.$$

The transform-domain compensation matrices for the 16 4×4 blocks are identical and are obtained as follows:

$$\Phi = \begin{bmatrix} 0 & 7b_2 & 0 & b_2 \\ 7c_2 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 \\ c_2 & 0 & 0 & 0 \end{bmatrix}$$

As it turns out, the frequency-domain compensation matrices for the plane prediction can be obtained at a minimal computational cost. In the pixel domain, apart from the derivation of a, b , and c , every position requires two multiplications. Hence, performing the compensation in the frequency domain reduces operations due to the elimination of the forward and inverse transforms (both DCT and Hadamard), and due to the sparseness of the frequency-domain compensation matrices. The calculation of the compensation in the frequency domain should have no impact on the quality when compared to the PDC architecture. However, since the downscaling after the inverse transform is performed in the PDC architecture, but not in the TDC architecture, this could result in small quality differences.

6 Implementation Results

In this section, we describe the results for the software implementation of the transcoding architectures on H.264/AVC bitstreams. The different transcoding architectures are tested using the Container and Stefan sequences, both in CIF resolution. The objective quality and the bitrate of the transcoded bitstreams are compared with the results obtained through re-encoding using a cascaded decoder-encoder pair. The bitstreams were encoded using the JVT reference software (Joint Model 9.8), restricted to the intra 16×16 modes only. The bitstreams were then transcoded from the initial QP_1 to a higher QP_2 ($\Delta QP = QP_2 - QP_1$), using the four architectures.

The results for the Stefan sequence are depicted in Table 3 ($\Delta QP = 5$) and Table 4 ($\Delta QP = 10$). Here, the luma PSNR (dB) and the bitrate (Mbps) of the original sequence at QP_1 are presented, along with the luma PSNR (dB) and the bitrate of the adapted bitstreams at QP_2 . The resulting bitrates are given in percentage of the bitrate of the original sequence.

For medium to high QPs, the PSNR difference between the fast transcoding architectures PDC and TDC, and the cascaded decoder-encoder pair remains limited to 0.5 to 1 dB. For very low QPs (high bitrates), this difference is larger.

Table 3. PSNR [dB] and bitrate [Mbps] results (Stefan, $\Delta QP = 5$)

QP_1	QP_2	Original		Cascade		Open-loop		PDC		TDC	
		PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate
4	9	57.1	17.14	50.5	84.5%	42.7	82.7%	45.7	84.8%	44.2	82.9%
10	15	51.2	12.66	45.2	80.8%	38.8	79.8%	42.2	81.0%	42.0	80.2%
16	21	46.4	9.02	40.2	77.3%	34.2	76.8%	38.7	77.7%	38.7	77.3%
22	27	41.4	6.20	35.1	71.9%	29.2	71.3%	34.3	72.3%	34.3	72.0%
28	33	36.4	4.03	30.2	63.3%	24.3	63.3%	29.5	64.3%	29.7	64.2%
34	39	31.5	2.44	25.7	53.4%	19.9	54.1%	25.0	55.6%	24.8	55.5%

Table 4. PSNR [dB] and bitrate [Mbps] results (Stefan, $\Delta QP = 10$)

QP_1	QP_2	Original		Cascade		Open-loop		PDC		TDC	
		PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate
4	14	57.1	17.14	47.2	70.9%	40.0	70.2%	43.9	71.3%	42.9	70.3%
10	20	51.2	12.66	42.0	66.8%	35.6	66.2%	40.1	67.0%	40.1	66.5%
16	26	46.4	9.02	36.8	61.3%	30.6	61.2%	35.8	61.9%	36.1	61.7%
22	32	41.4	6.20	31.8	53.8%	25.5	53.7%	31.0	54.3%	31.4	54.3%
28	38	36.4	4.03	27.1	43.7%	20.9	44.2%	26.4	45.1%	26.9	45.1%
34	44	31.5	2.44	23.1	33.1%	16.9	34.9%	22.2	36.2%	22.5	36.1%

Table 5. PSNR [dB] and bitrate [Mbps] results (Container, $\Delta QP = 5$)

QP_1	QP_2	Original		Cascade		Open-loop		PDC		TDC	
		PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate
4	9	57.1	13.95	50.5	81.5%	42.3	78.9%	45.2	81.8%	44.2	79.0%
10	15	51.2	9.78	45.1	73.9%	38.2	72.1%	42.2	74.0%	41.9	72.4%
16	21	46.2	6.58	40.1	66.6%	33.4	65.5%	38.7	66.8%	38.7	66.0%
22	27	41.3	4.11	35.7	60.4%	28.8	60.2%	34.6	61.5%	34.3	61.0%
28	33	36.8	2.39	31.9	55.4%	23.4	57.0%	30.9	58.6%	31.0	58.5%
34	39	33.0	1.33	28.2	51.8%	19.2	53.1%	27.0	55.3%	26.1	55.2%

Table 6. PSNR [dB] and bitrate [Mbps] results (Container, $\Delta QP = 10$)

QP_1	QP_2	Original		Cascade		Open-loop		PDC		TDC	
		PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate	PSNR	Bitrate
4	14	57.1	13.95	47.1	65.0%	39.5	63.5%	43.2	65.3%	43.0	63.5%
10	20	51.2	9.78	41.8	56.2%	34.4	54.8%	39.9	56.5%	39.7	55.1%
16	26	46.2	6.58	37.2	48.1%	29.8	47.8%	35.9	48.7%	36.1	48.3%
22	32	41.3	4.11	33.2	42.1%	25.1	43.0%	32.0	44.0%	32.3	43.9%
28	38	36.8	2.39	29.4	37.0%	19.1	39.5%	28.1	41.0%	28.8	41.0%
34	44	33.0	1.33	25.9	32.7%	14.4	35.1%	24.5	37.2%	24.7	37.2%

As mentioned in Sect. 4, this is caused by the non-linearity of the division after the inverse transform and the divisions in the DC and plane prediction modes. For higher PSNR values, the resulting defects become more distinct. Nonetheless, from the results it is clear that both compensation methods strongly outperform the traditional open-loop architecture at a negligible cost in bitrate reduction. Note that the PSNR values for bitstreams transcoded with TDC and PDC architectures are only slightly different. The results for the Container sequence are presented in Table 5 ($\Delta QP = 5$) and Table 6 ($\Delta QP = 10$), and are similar to the results obtained for the Stefan sequence.

The rate-distortion curves for the four architectures are shown in Fig. 6 ($\Delta QP = 5$) and Fig. 7 ($\Delta QP = 10$). These show that the rate-distortion performance decreases slightly for all four transcoder architectures when ΔQP is increased.

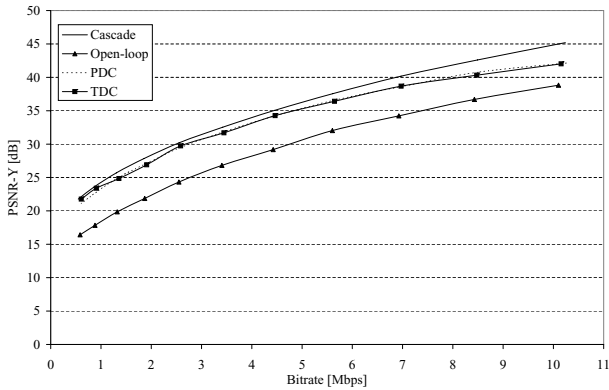


Fig. 6. Rate-distortion performance for $\Delta QP = 5$ (Stefan sequence)

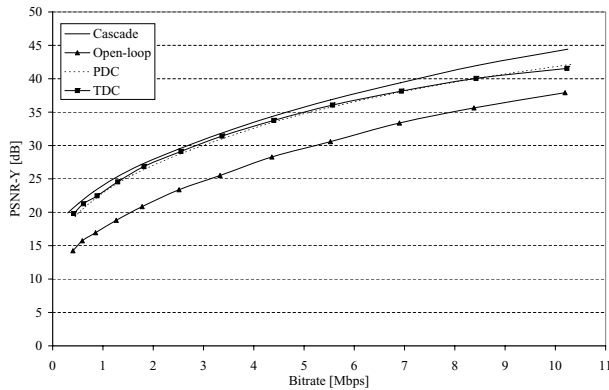


Fig. 7. Rate-distortion performance for $\Delta QP = 10$ (Stefan sequence)

7 Conclusions

In this paper, requantization techniques for H.264/AVC bitstreams were discussed, focusing on the intra 16×16 prediction. Two novel architectures were presented that solve the problem of drift propagation, as encountered for the more traditional open-loop requantization transcoder. Implementation results show that both architectures perform approximately equally well, and are able to approach the visual quality of a cascade of a decoder and a full-search encoder within 0.5 to 1 dB for medium to high quantization parameters. Because of the low computational complexity of the proposed architectures, they are highly suitable for on-the-fly rate reduction operations.

Acknowledgements

The research activities that have been described in this paper were funded by Ghent University, the Interdisciplinary Institute for Broadband Technology (IBBT), the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), the Belgian Federal Science Policy Office (BFSPO), and the European Union.

References

1. ITU-T and ISO/IEC JTC 1: Advanced video coding for generic audiovisual services. ITU-T Rec. H.264 and ISO/IEC 14496-10 AVC (2003)
2. Xin, J., Lin, C.W., Sun, M.T.: Digital Video Transcoding. *Proceedings of the IEEE* **93** (2005) 84–97
3. Vetro, A., Christopoulos, C., Sun, H.: Video Transcoding Architectures and Techniques: an Overview. *IEEE Signal Image Processing* (2003) 18–29
4. Sun, H., Kwok, W., Zdepski, J.W.: Architectures for MPEG compressed bitstream scaling. *IEEE Transactions on Circuits and Systems for Video Technology* **6** (1996) 191–199
5. Werner, O.: Requantization for Transcoding of MPEG-2 Intraframes. *IEEE Transactions on Image Processing* **8** (1999) 179–191
6. Richardson, I.E.G.: H.264 and MPEG-4 Video Compression: Video Coding for Next-generation Multimedia. John Wiley & Sons (2003)
7. Malvar, H., Hallapuro, A., Karczewicz, M., Kerofsky, L.: Low-Complexity Transform and Quantization in H.264/AVC. *IEEE Transactions on Circuits and Systems for Video Technology* **13** (2003) 598–603
8. Wiegand, T., Sullivan, G., Bjøntegaard, G., Luthra, A.: Overview of the H.264/AVC Video Coding Standard. *IEEE Transactions on Circuits and Systems for Video Technology* **13** (2003) 560–576

Motion-Compensated Deinterlacing Using Edge Information

Taeuk Jeong and Chulhee Lee

Dept. Electrical and Electronic Eng., Yonsei Univ.,
134 Shinchon-Dong, Seodaemun-Ku, 120-749 Seoul, Korea
Chulhee@yonsei.ac.kr

Abstract. In this paper, we propose a new deinterlacing method using motion estimation and compensation of edge regions. Although motion-compensated deinterlacing methods provide significant results in interlaced-to-progressive conversion, they still have undesirable artifacts in fast moving areas and edge regions. The proposed method mitigates the problems by applying the edge region motion estimation and compensation with properly small search range. After filling the missing lines with the conventional spatial and temporal methods, motion estimation and compensation is applied to the predefined edge areas. Experimental results show that the proposed method produces noticeable improvement more than existing motion-compensated deinterlacing methods.

1 Introduction

Recent display technology often requires the interlaced-to-progressive conversion. For examples, traditional SDTV video sequences are in interlaced scanned format and deinterlacing is necessary to display those videos on the progressive scanned devices such as PDPs, LCDs, and multimedia PCs. Deinterlacing is also required for frame rate conversions.

Among various deinterlacing methods, intrafield methods employ various vertical interpolations. They have been widely used in many applications due to the relatively low computational complexity. However, they yield undesirable artifacts in motion and edge areas. The directional spatial interpolation methods have been proposed to preserve edges [1-4]. Non-linear filtering methods using median filters have also been proposed with various spatial medians and vertical-temporal filters [5].

Motion adaptive methods [6-8] first determine motion information using pixel difference between adjacent fields. Then, deinterlacing is performed by using either spatial interpolation for stationary areas or temporal interpolation for motion areas. These methods provide improved picture quality for video sequences which contain moving objects within still background.

Motion-compensated deinterlacing algorithms [9-12] estimate the motion between adjacent fields and fill in the missing lines using motion information. Although they produce the best performance among various deinterlacing techniques, they may suffer from inaccurate estimation of motion vectors, which produce undesired results in fast motion. The DIMC (directional interpolation and motion-compensated)

deinterlacing method [12] uses directional interpolation and motion compensation. Motion-compensated adaptive method [13] has been proposed to mitigate the spurious errors from block-wise processing.

In motion-compensated deinterlacing, key degradation factors are the inaccurate estimation of motion vectors and relatively poor performance in edge areas. In order to address these problems, we propose a new motion-compensated deinterlacing method that uses edge information. The proposed edge-dependent motion compensation (EMC) deinterlacing method imposes a limit on the magnitude of motion vectors for edge regions.

2 Motion-Compensated Deinterlacing

In this section, we review briefly deinterlacing algorithms based on motion compensation and discuss some problems that may occur in motion-compensated deinterlacing methods such as the DIMC (directional interpolation and motion-compensated) deinterlacing algorithm. Fig.1 shows the block diagram of the DIMC method. Missing lines are filled by applying either directional intrafield interpolation or motion-compensated deinterlacing results based on a certain selection rule. Motion estimation is carried out between the same parity fields in order to obtain high quality since motion compensation between the same parity fields has inherent advantages [14].

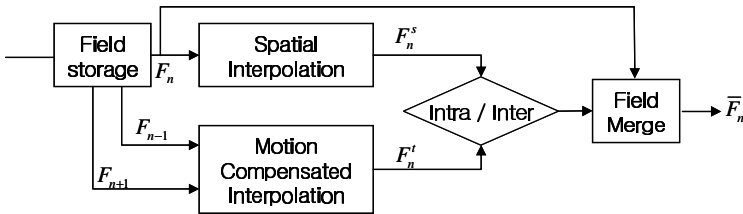


Fig. 1. Block diagram of the motion compensation algorithm

Motion vectors for macro blocks are determined after full search with a large search range (32x32). Since motion vectors are obtained in a half-pel resolution, the deinterlacing algorithm requires a long processing time. Furthermore, incorrect motion estimation results in performance degradation. In that case, motion-compensated deinterlacing does not necessarily provide better performance than intrafield deinterlacing. Thus, a mechanism which checks the reliability of results obtained from motion estimation is required in motion-compensated deinterlacing.

DIMC algorithm [12] first checks the reliability of motion estimation by comparing the variance of the block with the MSE. Motion estimation is considered as “reliable” if the MSE is smaller than the block variance or predetermined threshold value. Motion compensation is applied only to the “reliable” blocks as follows:

$$\bar{F}_n(x) = \begin{cases} F_n^s(x) & \text{if } \sum_{x \in B} |F_n^s(x) - F_n^t(x)|^2 > Th_1 \\ F_n^t(x) & \text{otherwise} \end{cases} \quad (1)$$

where $F_n^s(x)$ is the result obtained by applying intrafield deinterlacing and $F_n^t(x)$ the result obtained by applying motion-compensated deinterlacing.

Although motion-compensated deinterlacing methods provide better performance in most cases, they still suffer from undesired results in some cases. In particular, the DIMC algorithm tends to produce poorer performance than spatial methods in fast moving areas, non-rigid types of motion, and in edge regions [12]. In order to address the problems, we propose to impose a limit on motion estimation for moving areas and motion compensation for edge regions.

3 The Edge Motion Compensation Deinterlacing Method

In this section, we will describe the proposed edge-dependent motion compensation deinterlacing method (EMC). As can be seen in the Fig. 2, the proposed method first fills the missing lines of the current field F_n by selecting one of intrafield interpolation and conventional motion-compensated interfield interpolation. The motion estimation is applied to a predefined small search range between the two same parity fields: the previous field F_{n-1} and the next field F_{n+1} . Finally, the proposed method repeats motion-compensated deinterlacing only for edge areas.

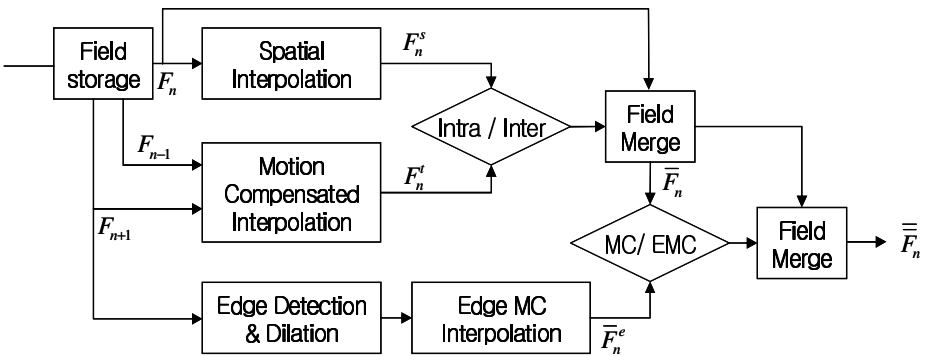


Fig. 2. Block diagram of the proposed edge MC algorithm

3.1 Limit on Motion Estimation

The correlation between the adjacent fields is small when there is a large amount of motion. In this case, motion estimation is not always reliable even if the large search range is considered, and motion-compensated deinterlacing may lead to performance degradation.

Another factor that affects the performance of motion-compensated deinterlacing is the magnitude of motion vectors. If the vertical length of a motion vector is odd, then a new pixel value is obtained from two interpolated pixels of the previous and next fields, as can be seen in Fig. 3(a), which tends to produce poor results. If the vertical length of a motion vector is even, we can expect better quality since the new pixel is calculated using two original pixels from the adjacent fields as can be seen in Fig. 3(b).

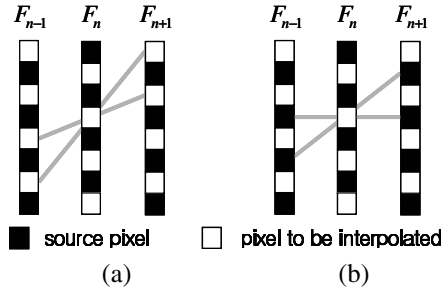


Fig. 3. Effects of the vertical length of a motion vector in motion (a) interpolated pixels will be used (b) original pixels will be used

Based on these observations, if the magnitude of motion vectors is small, it would produce better results for fast motion areas. Furthermore, if the vertical length is even, it may improve the performance of motion-compensated deinterlacing. In this paper, the search range for motion vectors is set as follows:

$$MV = \left\{ (v_x, v_y) : \left| \frac{1}{2} v_x \right| < 5 \text{ and } \left| \frac{1}{2} v_y \right| = 0, 2, 4 \right\}. \tag{2}$$

3.2 Motion-Compensated Deinterlacing Using Edge Information

In the proposed deinterlacing method, additional motion estimation is performed for edge areas within a limited search range between F_{n-1} and F_{n+1} . First, an edge image F_{n+1}^e of the next field F_{n+1} is obtained by conventional highpass filtering such as Sobel and Prewitt operators. Using a threshold value Th_m , an edge mask image F_{n+1}^{em} is generated as follows:

$$F_{n+1}^{em}(x) = \begin{cases} 1 & \text{if } F_{n+1}^e(x) > Th_m \\ 0 & \text{otherwise} \end{cases}. \tag{3}$$

Fig. 4a shows an edge image obtained using Sobel filter, Fig. 3b the corresponding mask image, and Fig. 3c an extended edge image obtained using edge dilation operation. The edge dilation operation is illustrated in Fig. 5. In the edge dilation operation, broken edges in the mask image are filled.

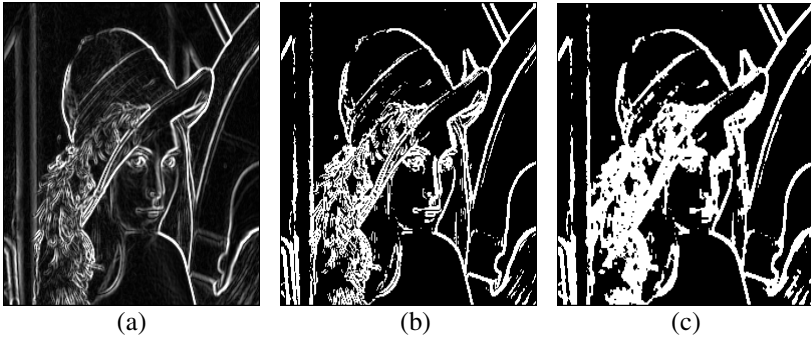


Fig. 4. (a) an image obtained using Sobel filter (b) a mask image (c) an extended edge image

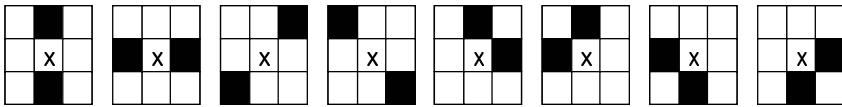


Fig. 5. Eight cases for the edge dilation operation

If $\sum_{x \in B} F_{n+1}^{em}(x) > Th_e$ for a macro block B in the next field, then the macro block is viewed as edge areas and new motion-compensated deinterlacing is performed for the macro block. The new motion vector v^e is determined as follows:

$$v^e = \arg \min_{v \in MV} \sum_{x \in B} F_{n+1}^{em}(x) | F_{n+1}(x) - F_{n-1}(x+v) |^2 \tag{4}$$

where the MV is the set of the predefined limited search range for motion vectors. The corresponding new pixel is obtained as follows:

$$\bar{F}_n^e(x+v_2^e+dmv) = \frac{1}{2} [F_{n-1}(x+v^e+dmv) + F_{n+1}(x+dmv)] \tag{5}$$

where $v_2^e = \frac{1}{2}v^e$ is a motion vector between next and current fields and dmv is displacement vector. Then, a reliability test, which is similar to that of DIMC, is employed. In the proposed method, the additional edge-dependent motion-compensated deinterlacing is applied only to the edge regions in the block which satisfy the following conditions:

$$\bar{\bar{F}}_n(x) = \begin{cases} \bar{F}_n(x) & \text{if } \sum_{x \in B} F_{n+1}^{em}(x) | \bar{F}_n(x) - \bar{F}_n^e(x) |^2 > Th_2 \\ \bar{F}_n^e(x) & \text{otherwise} \end{cases} \tag{6}$$

where $\bar{F}_n(x)$ is obtained using (1) and $\bar{F}_n^e(x)$ is obtained using (5). In other words, the proposed method applies edge-dependent motion-compensated deinterlacing to edge areas.

4 Experimental Results

Experiments were conducted in order to evaluate the performance of the proposed edge-dependent deinterlacing algorithm (EMC). The proposed method employs two spatial interpolation methods along with the proposed edge-dependent deinterlacing algorithm: the ELA method (ELA-EMC) and the directional interpolation with vertical interpolation function of the sampled 6-tap filter (DI6-EMC) of the Hamming windowed sinc function. An intrafield deinterlacing using the cubic spline interpolation method (CBI) [12], ELA [1] and a motion-compensated deinterlacing algorithm (DIMC [12]) are used for comparison.

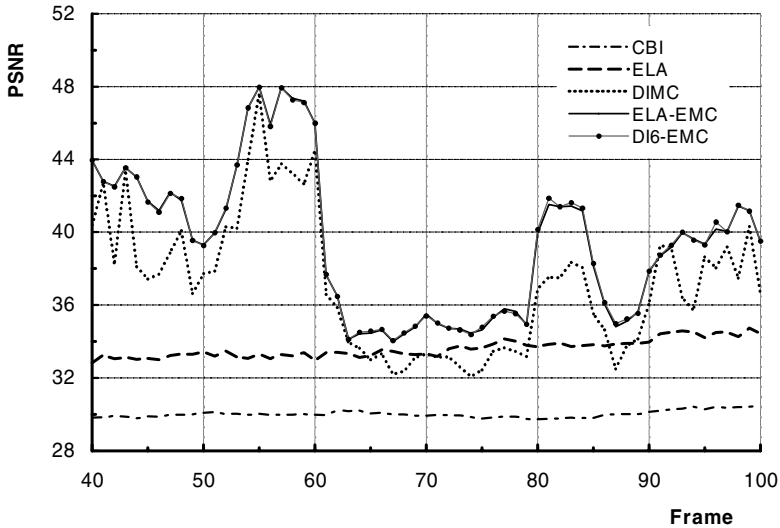


Fig. 6. PSNRs of the five methods for the Foreman sequence. CBI: cubic spline interpolation ELA: edge-based line average, DIMC: directional interpolation and motion compensation, DI6-EMC: the proposed edge MC with vertical interpolation function of 6-tap, ELA-SMC: the proposed edge MC with the spatial method of the ELA.

In the experiment, the threshold values Th_1 , Th_m , Th_e , and Th_2 are set to $w \times h \times 400$, 60, $w \times h \times 0.6$, and $400 \times \sum_{x \in B} F_{n+1}^{em}$, respectively. It is noted that w and h are a horizontal and vertical resolution of the macro block with $w=16$ and $h=8$. PSNRs between original and deinterlaced signals are calculated and used as performance criteria.

Fig. 6 shows the PSNRs of the five deinterlacing methods when they are applied to the Foreman sequence. The proposed algorithms (ELA-EMC and DI6-EMC) outperform the other methods. It is noted that DIMC provides a lower PSNR than ELA in the fast motion interval (between the 64th and 80th frames) while the overall performance of DIMC is superior to that of ELA.

Fig. 7 shows the difference images between the original image and the deinterlaced images of the 75th frame of the Foreman sequence. As can be seen in Fig. 7, the proposed methods show better performance in edge regions than the existing deinterlacing algorithms.

Table 1 shows the average PSNRs for eight test video sequences: four CIF format sequences (Coastguard, Mother & Daughter, Silent, Singer) and four QCIF format sequence (Container, Foreman, Mobile & Calendar, Table). The proposed methods outperforms the existing motion-compensated deinterlacing method (DIMC) by about 1.1 dB in PSNR.

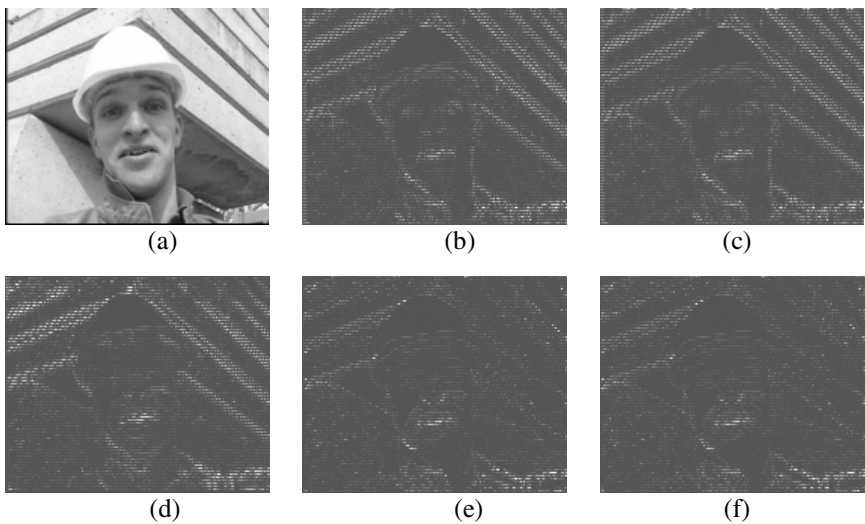


Fig. 7. Difference images of the five methods using the Foreman sequence (75th frame) (a) Original image (b) CBI (c) ELA (d) DIMC (e) ELA-EMC (f) DI6-EMC

Table 1. Average PSNRs for each video sequence

	CBI	ELA	DIMC	ELAEMC	DI6EMC
coastguard_cif	28.45	27.89	29.91	30.64	30.80
mother_daughter_cif	38.58	38.63	40.46	41.53	41.36
silent_cif	32.76	33.92	36.66	37.61	37.65
singer_cif	32.72	33.65	37.09	37.87	38.00
container_qcif	25.74	26.63	27.63	29.38	29.37
foreman_qcif	30.09	32.72	36.18	38.14	38.25
mobile_qcif	21.95	22.74	23.52	25.63	25.87
table_qcif	24.99	25.98	26.63	27.16	27.35
average	29.41	30.27	32.26	33.49	33.58

5 Conclusions

In this paper, we proposed a new deinterlacing algorithm that uses edge-dependent motion compensation. The proposed method first employs the conventional motion-compensated deinterlacing methods with a small search range. Then, the proposed method performs additional motion-compensated deinterlacing for edge areas. Experimental results show that the proposed algorithm provides noticeable performance improvement. In particular, the proposed method provides improved picture quality in edge regions.

Acknowledgments

This research was supported by the MIC(Ministry of Information and Communication), Korea, under the ITRC(Information Technology Research Center) support program supervised by the IITA(Institute of Information Technology Assessment). (IITA-2005-(C1090-0502-0027)).

References

1. T. Doyle and M. Looymans, "Progressive scan conversion using edge information," *Signal Processing of HDTV II*, L. Chairiglione, Ed. Amsterdam, The Netherlands: Elsevier, pp. 711-721, 1990.
2. C. J. Kuo, C. Liao, and C. C. Lin, "Adaptive interpolation technique for scanning rate conversion," *IEEE Trans. Circuits and systems for video technology*, vol. 6, no. 3, pp. 317 - 321, Jun. 1996.
3. H. Y. Lee, J. W. Park, T. M. Bae, S. U. Choi, and Y. H. Ha, "Adaptive scan rate up-conversion system based on human visual characteristics," *IEEE Trans. Consumer Electron.*, vol. 46, no. 4, pp. 999 - 1006, Nov. 2000.
4. H. Y and J. Jeong, "Direction-oriented interpolation and its application to de-interlacing," *IEEE Trans. Consumer Electronics*, vol. 48, no. 4, pp. 954 - 962, Nov. 2002.
5. J. Salo, Y. Nuevo, and V. Hameenaho, "Improving TV picture quality with linear-median type operation," *IEEE Trans. Consumer Electron.*, vol. 34, no. 3, pp. 373-379, Aug. 1988.
6. A. M. Bock, "Motion adaptive standards conversion between formats of similar field rates, signal processing," *Image Commun.*, vol. 6, no. 3, pp. 275-280, June 1994.
7. Renxiang Li, Bing Zeng and Ming L. Liou, "Reliable motion detection/compensation for interlaced sequences and its applications to deinterlacing," *IEEE Trans. Circuits and Systems for Video Technology*, vol. 10, no. 1, pp. 23-29, Feb. 2000.
8. D. Van De Ville, B. Rogge, W. Philips, and I. Lemabieu, "Deinterlacing using Fuzzy-Based Motion Detection," Knowledge-Based Intelligent Information Engineering systems, 1999. 3rd International Conference, 1999, pp. 263-267.
9. K. Sugiyama and H. Nakamura, "A Method of Deinterlacing with Motion-compensated Interpolation," *IEEE Trans. Consumer Electron.*, vol. 45, no. 3, pp. 611-616, Aug. 1999.
10. G. de Haan and E. B. Bellers, "Deinterlacing of video data," *IEEE Trans. Consumer Electron.*, vol. 43, pp. 819-825, Aug. 1997.

11. Y. Jung, B. T. Choi, Y. Park, and S. Ko, "An effective de-interlacing technique using motion-compensated interpolation," *IEEE Trans. Circuits and systems for video technology*, vol. 46, no. 1, pp. 460-466, Aug. 2000.
12. O. Kwon, K. Sohn and C. Lee, "Deinterlacing using Directional interpolation and motion compensation," *IEEE Trans. Consumer Electron.*, vol.49, no.1, pp. 198-203, 2003
13. S. Yang, Y-Y. Jung, Y. H. Lee, and R-H. Park, "Motion compensation assisted motion adaptive interlaced-to-progressive conversion," *IEEE Trans. Circuits and systems for video technology*, vol. 14, no. 9, Sep. 2004.
14. R. A. Beuker and I. A. Shah, "Analysis of interlaced video signals and its application," *IEEE Trans. Image Processing*, vol.3, no.7, pp. 501-552, Sep. 1994.

Video Enhancement for Underwater Exploration Using Forward Looking Sonar

Kio Kim, Nicola Neretti, and Nathan Intrator

Institute for Brain and Neural Systems
Brown University, Providence RI 02906, USA
kio@brown.edu

Abstract. The advances in robotics and imaging technologies have brought various imaging devices to the field of the unmanned exploration of new environments. Forward looking sonar is one of the newly emerging imaging methods employed in the exploration of underwater environments. While the video sequences produced by forward looking sonar systems are characterized by low signal-to-noise ratio, low resolution and limited range of sight, it is expected that video enhancement techniques will facilitate the interpretation of the video sequences. Since the video enhancement techniques for forward looking sonar video sequences are applicable to most of the forward looking sonar sequences, the development of such techniques is more crucial than developing techniques for optical camera video enhancement, where only specially produced video sequences can benefit the techniques. In this paper, we introduce a procedure to enhance forward looking sonar video sequences via incorporating the knowledge of the target object obtained in previously observed frames. The proposed procedure includes inter-frame registration, linearization of image intensity, and maximum a posteriori fusion of images in the video sequence. The performance of this procedure is verified by enhancing video sequences of Dual-frequency Identification Sonar (DIDSON), the market leading forward looking sonar system.

1 Introduction

Advances in robotics and imaging technologies have expanded the boundary of human activity and perception to those areas that have been out of our reach for a long time. The exploration of underwater environments is an example of successful applications of novel imaging and robotic technologies.

In the study of underwater environments, the use of forward looking sonar (FLS) systems is increasing thanks to the high frame rate, relatively high resolution, low power consumption and portability [1,2,3]. Forward looking sonar is a type of sonar that produces a 2D image by stacking 1D images produced by a 1D transducer array. Unlike in conventional sonar, the beam forming of FLS is spontaneously achieved without additional computation, so it can produce relatively high resolution images at a frame rate comparable to that of optical video cameras. There are several high performance FLS systems that are commercially available, and the use of such sonar systems is increasing these days [4,5].

Despite the merits of FLS systems, it has shortcomings when compared to optical cameras [6]. First, the angular resolution is relatively low, typically less than 100 pixels.

Second, the signal-to-noise ratio is still lower than that of optical cameras because of the nature of B-scan images.

In this paper, we present a procedure to enhance FLS video sequences by fusing information collected from different frames. This procedure can reduce noise and increase the spatial resolution of FLS video sequences.

2 Scope of Applicable Video Sequences

Video enhancement algorithms based on super-resolution techniques have been extensively studied [7,8,9]. Such algorithms are basically made up in three parts including i) the registration, ii) the transformation, and iii) the fusion of the images. For optical cameras, the scope of video sequences that can be processed by video enhancement algorithms is strictly limited by the requirements in each step of the procedure, while the FLS video sequences are free from such restrictions. Once an enhancement method for FLS is established, it can be applied to most of FLS video sequences. For this reason, the development of a good video enhancement method is more important in FLS than in conventional optical cameras.

For image registration, one needs to model the homographic relation between images based on the imaging geometry of the imaging device. For example, when a pinhole camera views a planar surface from different perspectives, a perspective transformation is sufficient to explain the homography of images. Most of enhancement methods for optical camera images use a perspective homography or similar homographies of lower hierarchy, such as affine homography. A projective homography requires, however, the target object to be a planar surface, or the camera undergoes only rotational motion without translational motion. For optical cameras, mostly those video sequences intentionally produced can satisfy this requirement. In contrast, FLS requires the target object to be on a planar surface from the image acquisition level—otherwise, the visibility of sonar is extremely narrowed, and the output images suffer significant vignetting. (See Fig. 1.) This property imposes a huge constraint to the variability of FLS images so that an affine transformation can explain most of FLS video sequences [6].

For the fusion of images, in order to combine multiple frames of optical camera video sequences, one needs a video sequence without any occlusion in it. Or, when an occluded area exists in a scene, one needs to add extra steps for segmenting and ex-

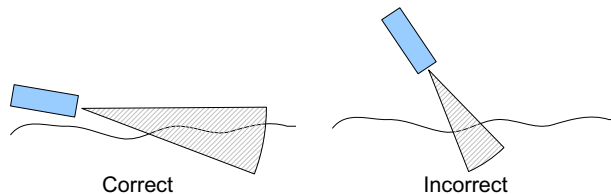


Fig. 1. Correct (left), and incorrect use (right) of a forward looking sonar system. When a FLS device is incorrectly used as describe above, the visible area in the produced image is significantly restricted.

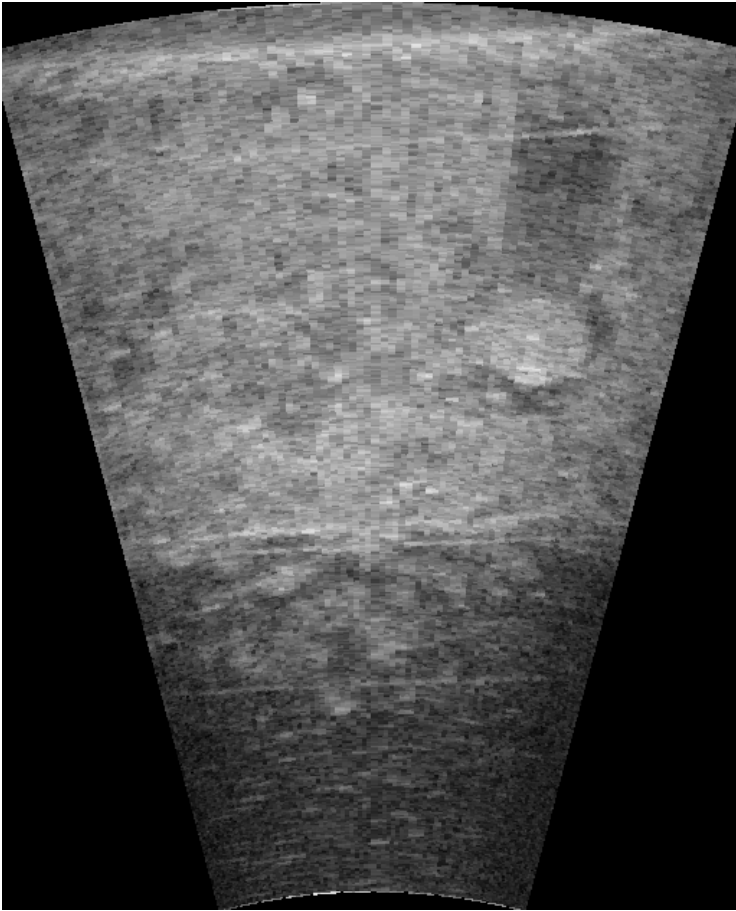


Fig. 2. Occlusion by a cylindrical object in a scene of a forward looking sonar video sequence. The occluded area appears darker than its usual appearance, instead of being replaced by the shape of the occluding object as in optical camera images.

cluding the occlusion. Another merit of FLS video sequences is that the fusion method can tolerate occlusions as far as the occlusion is static. In FLS images, an area turns dark when it is occluded, and recovers its original brightness as soon as it escapes the occlusion. This is simply a change of illumination, which is much easier to handle than an occlusion exclusion problem in optical camera video enhancement. (See Fig. 2.)

3 Methodology

The proposed procedure is largely made up of the following steps: i) separation of illumination profile, ii) inter-frame image registration, iii) linearization of brightness, and iv) maximum a posteriori (MAP) fusion of images.

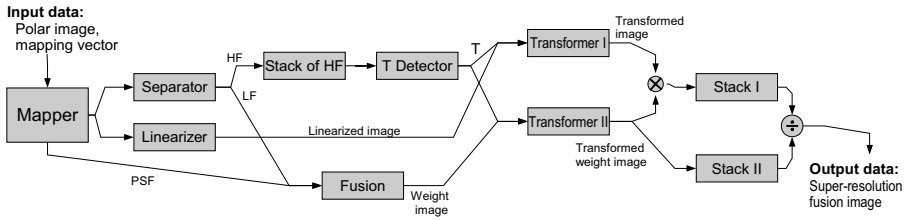


Fig. 3. Block diagram of the super-resolution image fusion process

The flow of data in the procedure is depicted in Fig. 3. The detailed description of each step in the procedure is presented in the following subsections.

In the step i), an image is separated into the high frequency part and the low frequency part. In the step ii), the registration is performed both between two neighboring frames and between frames apart. The parameters of registration found therein are combined optimally to minimize the accumulation of registration errors. In the step iii), image intensity of the images is linearized for the maximum a posteriori fusion of the sequence in the following step. In the step iv), the previously observed frames are fused into one image to best render the frame displayed at the moment. The detailed procedures are described below.

3.1 Retinex Separator

Unlike in optical camera images, the illumination condition of FLS varies significantly within a sequence, and also within a frame, because the illumination depends merely on the ultrasound beam incident from the device itself. A slight change of the grazing angle of the FLS device and the curvature of the target surface can bring a variation of the illumination condition, which eventually makes the registration and fusion difficult. For this reason, one needs to separate the illumination profile from the reflectance profile of the target object.

Land has modeled an illumination process as homomorphic filtering [10], and the consequent researches disclosed algorithms to separate the illumination profile and the reflectance profile of an image in that regard [11,12]. For FLS images, a simple homomorphic filtering of the image is sufficient for the separation, say,

$$HF(\mathbf{x}) = I(\mathbf{x})/LF(\mathbf{x}), \quad (1)$$

where $I(\mathbf{x})$, $LF(\mathbf{x})$, and $HF(\mathbf{x})$ represent the intensity values at the position \mathbf{x} in the original image, the low frequency part, and the high frequency part, respectively. The low frequency part is calculated by low-pass filtering the image. We consider that LF and HF are the illumination profile and the reflectance profile, respectively in this paper.

3.2 Inter-frame Registration

In previous work of the authors, it has been discussed that the cross-correlation based feature matching outperforms the conventional feature matching algorithms, particularly for detecting correspondence of images with low resolution [6]. With the outliers

```

(1) anchor_frame = 1
(2) for current_frame = 2:end_of_sequence
(3)   Register current_frame with anchor_frame.
(4)   if anchor_frame != current_frame-1,
(5)     Register current_frame with current_frame-1.
(6)   end if
(7)   if the registration above fails,
(8)     Optimize valid transformation parameters
       between anchor_frame and current_frame-1.
(9)     reset anchor_frame = current_frame
(10)  else if current_frame-anchor_frame == predefined_number,
(11)    Optimize transformation parameters until
       between anchor_frame and current_frame.
(12)    Reset anchor_frame = current_frame + 1.
(13)  end if
(14)end for

```

Fig. 4. The algorithm for inter-frame registration

removed via an appropriate algorithm such as RANSAC (Random Sampling Consensus) or LMedS (Least Median of Squares), these feature point pairs serve to register images with subpixel accuracy.

However, even when a subpixel accuracy registration between two frames is obtained, there still remains a concern about the accumulated registration error in registering multiple frames in a video sequence. Further more, the registration of FLS images is based on an affine approximation of the homography [6] instead of the exact geometrical model, the accumulation of registration error can lead to even more significant errors in the registration. A fine tuning of the registration parameters is performed in order to address this problem.

The ideal condition for the least registration error is when all the frames in the sequence are consistently registered with one another. In most of cases, however, the camera is in motion and the view of the targeted area can evolve during the image acquisition period, so most of the frames can be registered with only a few of their neighboring frames.

The proposed algorithm determines how many neighboring frames can be registered with one frame (called an ‘anchor frame’). The first frame of the sequence of interest is set to be the anchor frame (Step (1) of Fig. 4), and the following frames are registered with the anchor frame (Step (3)), as well as their previous frame (Step (5)), until any of the registrations fails (Step (7)). When the size of the registrable section for the anchor frame is determined (Step (7)), one calculates the optimal set of transformation parameters that explains the homographies of all the frames in the section, with the minimal error (Step (8)). After this optimization step, it moves on to the remaining part of the sequence with the anchor frame reset to be the first first frame of the remaining part, until it reaches the end of the sequence (Step (9)). The structure of the algorithm is described in Fig. 4.

The maximum size of a registrable section has been limited in order to prevent the excessive dimensionality in optimization, to attain the desired latency of process under

the allowed computational power, and also to meet the desired performance of the image enhancement (Step (10) of Fig. 4). The optimization of transformation parameters is done by minimizing the energy function defined as

$$\begin{aligned}
 & E(\mathbf{p}_{1,2}, \mathbf{p}_{2,3}, \dots, \mathbf{p}_{n-1,n}) \\
 &= \sum_{k=1}^{n_{1,2}} |FP_{1,2,2}^k - T(\mathbf{p}_{1,2})FP_{1,2,1}^k|^2 \\
 &+ \sum_{j=3}^n \left\{ \sum_{k=1}^{n_{1,j}} |FP_{1,j,j}^k - T(\mathbf{p}_{1,k})FP_{1,j,1}^k|^2 \right. \\
 &\quad \left. + \sum_{k=1}^{n_{j-1,j}} |FP_{j-1,j,j}^k - T(\mathbf{p}_{j-1,j})FP_{j-1,j,j-1}^k|^2 \right\}, \tag{2}
 \end{aligned}$$

where n is the number of frames in the registrable section of the anchor frame, and $n_{p,q}$ is the number of inlier feature point pairs in the registration of the p -th and the q -th frames. $FP_{p,q,r}^k$ is the position vector of the k -th inlier feature point pair of the registration of the p -th and the q -th frames found in the r -th frame, and $T(\mathbf{p}_{p,q})$ is the transformation operator defined by the registration parameter $\mathbf{p}_{p,q}$. For any two non-consecutive frame numbers i and j , $\mathbf{p}_{i,j}$ is obtained by combining all the transformation parameters of the consecutive frames between the i -th and the j -th frames, say, $\mathbf{p}_{i,i+1}, \dots, \mathbf{p}_{j-2,j-1}, \mathbf{p}_{j-1,j}$.

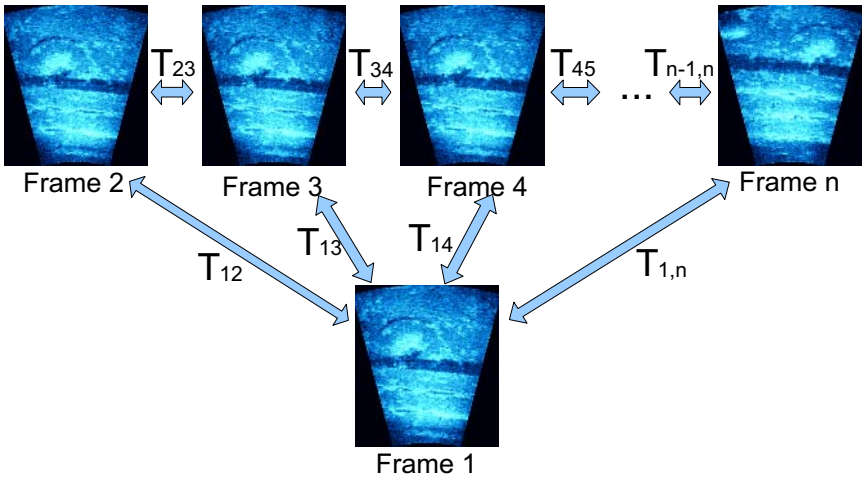


Fig. 5. Paring of the frames in an inter-frame registration of a section of frames. The first frame of a section, or the anchor frame, is registered with all other frames, and all the frames in this section are registered with its neighbors in the section as well.

3.3 Linearization of Image Intensity

When the strength of noise is comparable to the strength of signals as in ultrasound B-scan images, the noise structure is better explained by Rician statistics than Gaussian [13]. The Rician noise in general has non-zero mean, and the mean value of this additive noise is a function of the signal intensity, which in effect distorts the linearity of the image intensity. In addition to the distortion of linearity by the Rician noise, one has to consider the response property of the sensors in the imaging device, which might have been tuned to the precision that is sufficient only for visualization. Since the MAP fusion that will be described in the following subsection requires higher linearity of the sensor response, an additional tuning has to be performed.

For example, for DIDSON images, the following linearizing function significantly improves the quality of the fusion:

$$I'(\mathbf{x}) = \begin{cases} \{I(\mathbf{x}) - \mu\}^2, & \text{if } I(\mathbf{x}) > \mu \\ 0, & \text{otherwise} \end{cases} \quad (3)$$

where $I(\mathbf{x})$, μ , and $I'(\mathbf{x})$ represent the intensity at the position vector \mathbf{x} , the average background noise level, and the linearized intensity at \mathbf{x} , respectively.

3.4 Approximated MAP Image Fusion

Once the inter-frame registration and the linearization steps are complete, the frames are ready to undergo the final step of the procedure—the fusion step. Kim et al. have shown that the maximum a posteriori estimation of an image based on a set of low quality observation images of the image can be approximate by a weighted fusion of the low quality images [14]. This implies that one can perform the MAP image fusion without an iterative calculation that many of the super-resolution algorithms require. In addition, the method therein provides robustness under the inhomogeneous illumination condition which is occasional in FLS images.

The enhancement of a frame in a FLS video sequence is attained by fusing a predefined number of frames into the desired perspective using the MAP fusion as described in [14]; when N low resolution images β_1, \dots, β_N are fused,

$$\bar{\theta} \simeq \left(\sum_{i=1}^N W_i + v_0 V_0^{-1} \right)^{-1} \left(\sum_{i=1}^N W_i M_i \beta_i \right), \quad (4)$$

where $\bar{\theta}$, W_i , M_i and $v_0 V_0^{-1}$ represent the calculated MAP fusion image, the i -th reliability matrix, the i -th up-sampling matrix, and the regularization factor, respectively. The up-sampling matrix M_i is a n_{HR}^2 -by- n_{LR}^2 matrix, where n_{HR}^2 and n_{LR}^2 are the number of pixels in the high resolution image and in the low resolution image, respectively. The reliability matrix W_i is a n_{HR}^2 -by- n_{HR}^2 matrix, which includes all the factors that affect the reliability of a pixel value, for example, illumination intensity, point spread function, etc. The regularization factor $v_0 V_0^{-1}$ is basically the inverse of the covariance matrix of pixel values normalized by v_0 , the generic variance of pixel values. Ideally it includes non-zero off-diagonal terms, but for the simplicity of calculation, it is approximated by a diagonal matrix.

4 Results

An experiment was performed on a video sequenced used in a ship hull inspection [3]. In the inspection, a dual-frequency identification sonar (DIDSON) system mounted on a remotely controlled vehicle recorded the images of the surface of a ship hull while the vehicle was manipulated to move around on the ship hull surface.

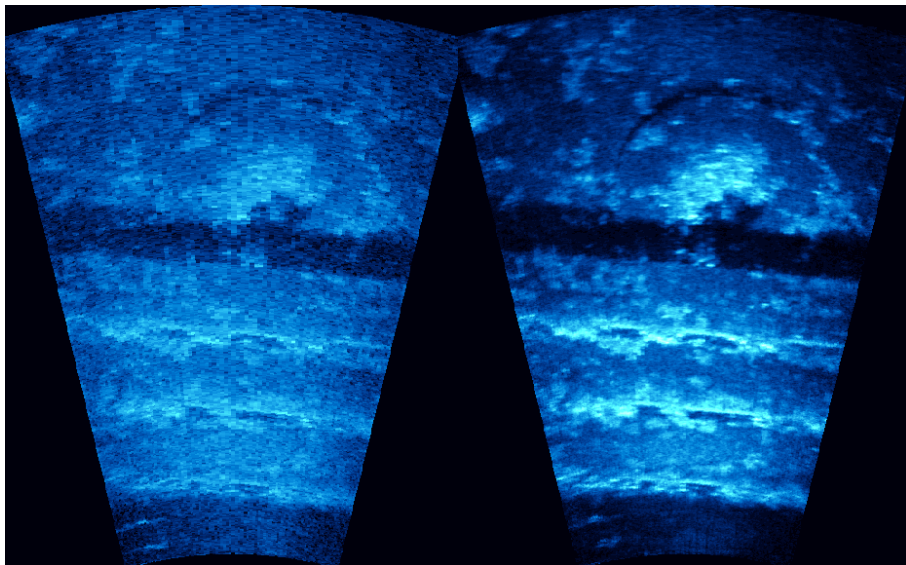


Fig. 6. Comparison of a frame in the original (left) and the enhanced (right) sequences. The frame in the enhanced sequence is a fusion of 7 consecutive frames portrayed in the same perspective.

The resolution of the initial polar coordinates images is 96×512 , and the polar coordinate images are mapped to images of size 512×576 size in the Cartesian coordinate system. The size of the Cartesian coordinate image is approximately the size that the smallest pixel in the polar coordinates image can occupy at least one pixel in the Cartesian coordinate, and mostly more than one pixel. In this way, one pixel in the polar coordinates occupies from 1 pixel to up to 20 pixels in the Cartesian coordinates, due to the fixed field-of-view of a sensor and varying distance from the transducer to the observed area.

The suggested procedure has been applied to the video sequence. Figure 6 is the comparison of one of the frames in the original sequence, and the enhanced sequence, where up to 7 neighboring frames were fused to re-render a frame. In Fig. 6, one can verify that the fusion image (right) discloses crispier edges of the target object, than the original image (left). Also note that the surface texture that was difficult to identify in the original sequence can be easily identified in the enhanced sequence due to the reduced noise level and the increased solution.

5 Conclusion

In this paper, we presented a procedure to enhance a forward looking sonar video sequence. The procedure includes the separation of illumination profile, inter-frame registration, the linearization of the images, and non-iterative maximum a posteriori fusion of images for super-resolution.

Since the image acquisition method of FLS restricts the applicable target object to be on a planar surface, most of the FLS images can be registered using a simple affine homography. In addition, the occlusion problem, which often is an obstacle in processing optical video sequences, can be viewed simply as an illumination problem in FLS video sequences, which can be dealt with little trouble. This means there is no need to further consider the occlusion problem in FLS video sequences. For these reasons, video enhancement techniques for FLS in general are applicable to most of the FLS video sequences.

The proposed video enhancement procedure is largely made up of four steps including the separation of illumination profile, fine tuning of the registration parameters via inter-frame image registration, the linearization of brightness, and the maximum a posteriori (MAP) fusion of the images. All these steps are achievable with low computational power.

In future, further study for real time implementation of the proposed procedure is anticipated.

References

1. R. A. Moursund and T. J. Carlson and R. D. Peters. A fisheries application of a dual-frequency identification sonar acoustic camera. *ICES Journal of Marine Science*, 60(3):678–683, 2003.
2. Yunbo Xie, Andrew P. Gray, and Fiona J. Martens. Differentiating fish targets from non-fish targets using an imaging sonar and a conventional sonar: Dual frequency identification sonar (didson) versus split-beam sonar. In *Journal of the Acoustical Society of America*, volume 117, pages 2368–2368, 2005.
3. J. Vaganay, M. L. Elkins, S. Willcox, F. S. Hover, R. S. Damus, S. Desset, J. P. Morash, and V. C. Polidoro. Ship hull inspection by hull-relative navigation and control. In *Proceedings of Oceans '05 MTS/IEEE*, pages –, 2005.
4. <http://www.didson.com>.
5. <http://www.blueviewtech.com>.
6. K. Kim, N. Neretti, and N. Intrator. Mosaicing of acoustic camera images. *IEEE Proceedings—Radar, Sonar and Navigation*, 152(4):263–270, 2005.
7. M. Irani and S. Peleg. Motion analysis for image enhancement: Resolution, occlusion, and transparency. *Journal of Visual Communication and Image Representation*, 4(4):324–335, 1993.
8. S. Borman and R. Stevenson. Spatial resolution enhancement of low-resolution image sequences - a comprehensive review with directions for future research. Technical report, 1998.
9. Sung Cheol Park, Min Kyu Park, and Moon Gi Kang. Super-resolution image reconstruction: A technical overview. *IEEE Signal Processing Journal*, 20(3):21–36, 2003.
10. E. H. Land and J. J. McCann. Lightness and the retinex theory. *Journal of the Optical Society of America*, 61:1–11, 1971.
11. E. H. Land. An alternative technique for the computation of the designator in the retinex theory of color vision. *PNAS*, 83:3078–3080, 1986.

12. D. J. Jobson, Z. Rahman, and G. A. Woodell. Properties and performance of the center/surround retinex. *IEEE Transactions on Image Processing*, 6:451–462, 1997.
13. R. F. Wagner, S. W. Smith, J. M. Sandrik, and H. Lopez. Statistics of speckle in ultrasound b-scans. *IEEE Transactions on Sonics and Ultrasonics*, 30(3):156–163, 1983.
14. K. Kim, N. Neretti, and N. Intrator. Maximum a posteriori fusion method for super-resolution of images with varying reliability. pages –, 2006.

Optimal Parameters Selection for Non-parametric Image Registration Methods

Jorge Larrey-Ruiz and Juan Morales-Sánchez

Grupo de Teoría y Tratamiento de la Señal,
Departamento de las Tecnologías de la Información y las Comunicaciones,
Universidad Politécnica de Cartagena, Cartagena (Murcia) 30202, Spain
{jorge.larrey, juan.morales}@upct.es
<http://gtts.upct.es>

Abstract. Choosing the adequate registration and simulation parameters in non-parametric image registration methods is an open question. There is no agreement about which are the optimal values (if any) for these parameters, since they depend on the images to be registered. As a result, in the literature the parameters involved in the registration process are arbitrarily fixed by the authors. The present paper is intended to address this issue. A two-step method is proposed to obtain the optimal values of these parameters, in terms of achieving in a minimum number of iterations the best trade-off between similarity of the images and smoothness of the transformation. These optimal values minimize the joint energy functional defined in a variational framework. We focus on the specific formulation of diffusion and curvature registration, but the exposed methodology can be directly applied to other non-parametric registration schemes. The proposed method is validated over different registration scenarios.

1 Introduction

Image registration is the process of finding an optimal geometric transformation that aligns points in one view of an object with corresponding points in another view of the same object or a similar one. Particularly, in medical imaging there are many applications that demand for registration (e.g. medical image fusion, atlas matching, pathological diagnosis). During the past two decades, many methods have been proposed to set into correspondence monomodal or multimodal medical images, leading to a flourishing literature. For an overview on registration methods, we refer to e.g. [1], [2], and more particularly to [3], [4], [5], and references therein, for medical image registration.

In many applications, rigid registration (i.e., a registration based on rotations and translations) does not provide a sufficient solution. A non-linear (non-rigid) transformation is necessary to correct the local differences between the images. Non-rigid image registration can be either parametric or non-parametric. For parametric techniques (see e.g. [6], [7]), the transformation can be expanded in terms of some parameters of basis functions. The required transformation is a

minimizer of the distance measure in the space spanned by the basis functions. The minimizer can be obtained from some algebraic equations or by applying appropriate optimization tools.

For non-parametric techniques the registration is based on the regularized minimization of a distance measure. A regularizing term is used to circumvent ill-posedness and to privilege more likely solutions. It is the regularizer which distinguishes the existing methods: e.g. Elastic [8], Fluid [9], Diffusion [10], and Curvature [11] registration schemes. The main difference with respect to the parametric case, where one is looking for optimal parameters of an expansion of the transformation, is that we are now simply seeking a smooth transformation, without any parameter involved in representing it. Anyhow, non-parametric methods also require some parameters to control the evolution of the partial mappings towards the final result.

In contrast to many other ill-posed problems, where efficient strategies like the Generalized Cross Validation (GCV) approach [12] are available to automatically estimate the regularization parameters, for image registration such satisfactory strategies are yet missing [13], leading the authors of previous works to arbitrarily fix these parameters. The subsequent validation of the registration results is commonly performed by trained experts. In many cases, the value of the parameters does not have a big impact on the accuracy of the registration (i.e., a rough setting of these values is allowed). However, this setting may become more of a factor when the computed transformation is applied to clinical data, in which the matching is a more difficult task [14]. This paper is intended to address this problem by providing the guidelines on how to choose the registration and simulation parameters for non-parametric image registration methods, allowing for an optimal registration in terms of both similarity of the images and smoothness of the transformation, and minimizing the number of iterations of the registration algorithm.

The paper is organized as follows. We start out with the mathematical formulation of the general registration problem, introducing the regularization terms to be considered all over this work. This section is followed by the proposed methodology, which consists of two sequential steps. In section 4, we apply this new approach and prove the effectiveness of the exposed method on three realistic examples. Finally, the main issues presented throughout the paper are discussed.

2 Mathematical Framework

Given two images, a reference $I \equiv I(\mathbf{x})$ and a template $J \equiv J(\mathbf{x})$, with $\mathbf{x} \in \Phi \equiv]0, 1[^d$, the aim of image registration is to find a global and/or local transformation from J onto I in such a way that the transformed template matches the reference. Then the purpose of the registration is to determine a displacement field $\mathbf{v} \equiv \mathbf{v}(\mathbf{x})$ such that $J_{\mathbf{v}} \equiv J(\mathbf{x} - \mathbf{v}(\mathbf{x}))$ is similar to $I(\mathbf{x})$ in the geometrical sense. It turns out that this problem may be formulated in terms of a variational approach [15], [16], [17]. To this end we introduce the joint energy functional to be minimized

$$\mathcal{E}_{joint}[\mathbf{v}] = \mathcal{E}_{sim}[I, J; \mathbf{v}] + \lambda \mathcal{E}_{reg}[\mathbf{v}] , \tag{1}$$

where \mathcal{E}_{sim} represents a distance measure (external forces) and \mathcal{E}_{reg} determines the smoothness of the displacement field \mathbf{v} (internal constraints). The second term is unavoidable, because arbitrary transformations may lead to cracks, foldings, or other unwanted deformations. Therefore \mathcal{E}_{reg} can be seen as a regularizing term introduced to distinguish particular transformations which must be more likely than others. The resulting transformation should be a homeomorphism, i.e., a continuous bijection with a continuous inverse. The parameter λ is used to control the strength of the smoothness of the displacement vectors versus the similarity of the images.

Probably the most popular choice for the distance measure is provided by the so-called sum of squared differences (SSD):

$$\mathcal{E}_{sim}[I, J; \mathbf{v}] = \frac{1}{2} \int_{\Phi} (J_{\mathbf{v}} - I)^2 d\mathbf{x} . \tag{2}$$

For this measure to be successful, one has to assume that a monomodal registration is being performed (i.e., the intensities of the two given images are comparable).

In this paper we focus on two different smoothing terms, defined by the equations (3) and (4). Anyhow, the methodology exposed in this work can be directly applied to other non-parametric registration schemes (e.g. elastic registration).

$$\mathcal{E}_{reg}[\mathbf{v}] = \frac{1}{2} \sum_{l=1}^d \int_{\Phi} \|\nabla v_l\|^2 d\mathbf{x} , \tag{3}$$

$$\mathcal{E}_{reg}[\mathbf{v}] = \frac{1}{2} \sum_{l=1}^d \int_{\Phi} (\Delta v_l)^2 d\mathbf{x} , \tag{4}$$

where Δ is the two-dimensional Laplace operator.

The first technique (equation (3)) is known as *diffusion* registration [10]. It borrows the optical flow regularizer [18] and it is also related to Thirion’s so-called demons registration [19]. The reasons for this particular choice are twofold: it is designed to penalize oscillating deformations and it consequently leads to smooth displacement fields; and it permits a fast and efficient implementation.

The second technique (equation (4)) is known as *curvature* registration [11]. It is based on second order spatial derivatives, so its kernel contains affine linear transformations (i.e., an additional pre-registration step, unavoidable in the diffusion scheme, becomes redundant in this case). Curvature registration also allows for an efficient implementation.

According to the calculus of variations, the Gâteaux variation of the joint energy functional should be zero, i.e. a displacement field \mathbf{v} minimizing equation (1) necessarily should be a solution for the Euler-Lagrange equation

$$-\mathbf{f}(\mathbf{x}; \mathbf{v}) + \lambda \mathcal{A}[\mathbf{v}](\mathbf{x}) = \mathbf{0} \tag{5}$$

subject to appropriate boundary conditions. $\mathcal{A}[\mathbf{v}] \equiv \mathbf{A}\mathbf{v} \equiv -\Delta\mathbf{v}$ (for diffusion registration) or $\mathcal{A}[\mathbf{v}] \equiv \mathbf{A}\mathbf{v} \equiv \Delta^2\mathbf{v}$ (for curvature registration) is a partial differential operator related to the smoother \mathcal{E}_{reg} [10], [11]. The force field

$$\mathbf{f}(\mathbf{x}; \mathbf{v}) = (J_{\mathbf{v}} - I) \nabla J_{\mathbf{v}} \tag{6}$$

is used to drive the deformation. Changing the distance measure \mathcal{E}_{sim} results in a different force field. One common way to solve a non-linear partial differential equation like (5) is to introduce an artificial time t and to compute the steady-state solution of the time-dependent partial differential equation (i.e., $\partial_t\mathbf{v}(\mathbf{x}, t) = \mathbf{0}$) via a time-marching algorithm, letting the time derivative $\partial_t\mathbf{v}(\mathbf{x}, t)$ equal to the negative Gâteaux derivative of the registration energy functional [17]. In the diffusion case, the resulting expression is an inhomogeneous heat equation

$$\partial_t\mathbf{v}(\mathbf{x}, t) - \lambda \Delta\mathbf{v}(\mathbf{x}, t) = \mathbf{f}(\mathbf{x}, \mathbf{v}(\mathbf{x}, t)) \ . \tag{7}$$

For the curvature scheme, the resulting expression is the bi-harmonic equation

$$\partial_t\mathbf{v}(\mathbf{x}, t) + \lambda \Delta^2\mathbf{v}(\mathbf{x}, t) = \mathbf{f}(\mathbf{x}, \mathbf{v}(\mathbf{x}, t)) \ . \tag{8}$$

Equations (7) and (8) can be discretized in time and space, and then solved by employing the following semi-implicit iterative scheme

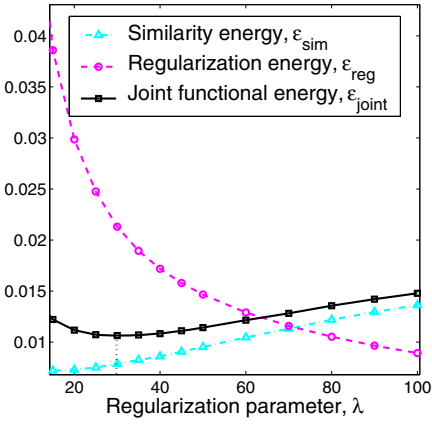
$$v_l^{(k+1)} = \left(\mathbf{I} - \tau\lambda\mathbf{A}\right)^{-1} \left(v_l^{(k)} + \tau f_l^{(k)}\right) \quad l = 1, \dots, d \ , \tag{9}$$

whose numerical solution allows for efficient implementations based on the discrete cosine transform (it can be deduced from [20]), or on an additive operator splitting (AOS) scheme (see [10]). Note that on the right hand side of equation (9), τ compromises between the current displacements and the current forces. In our implementation, the forces are scaled so that the time-step can be fixed to a value of $\tau = 1$.

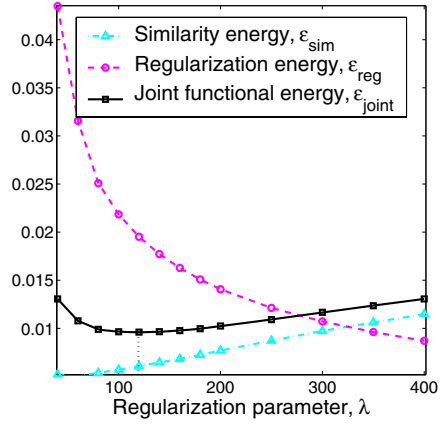
3 Proposed Methodology

The methodology exposed in this paper consists of two sequential steps:

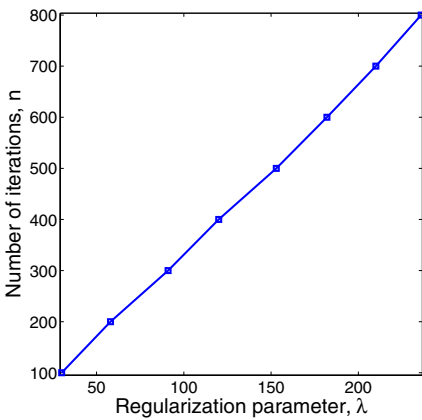
1. *Initial estimation of the parameters relation.* For a small number of iterations \hat{n} (typically between 100 and 200), the value of the regularization parameter $\hat{\lambda}$ that minimizes the joint energy functional (1) is obtained, as seen in figure Fig.1(a) (in this example, where a diffusion registration is performed, $\hat{n} = 100$ and $\hat{\lambda} = 30$). Note that due to the small value of \hat{n} , the computational load of this step is relatively light. In order to compute the registered template $J_{\mathbf{v}}$ and the displacement field \mathbf{v} for each value of λ , equations (6), (9) and an efficient DCT implementation of the algorithm (with suitable boundary conditions, see [21]) have been used. In order to compute the similarity and regularization energies, equations (2) and (3) have also been used. At this



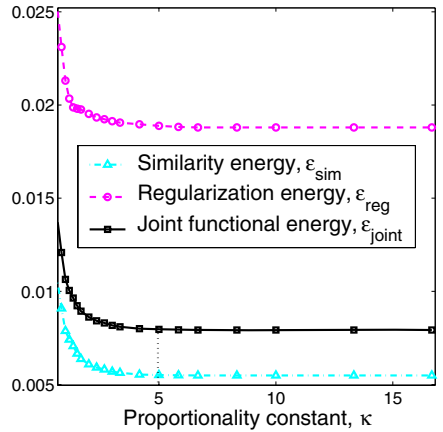
(a) Step 1. Joint functional energies ($\hat{n} = 100$).



(b) Joint functional energies ($\hat{n} = 400$).



(c) Regularization parameter $\hat{\lambda}$ versus number of iterations \hat{n} .



(d) Step 2. Joint functional energies ($\eta = 3.333$).

Fig. 1. Proposed Methodology

point, the quotient η between the number of iterations and the regularization parameter can be calculated as

$$\eta = \frac{\hat{n}}{\hat{\lambda}} . \tag{10}$$

The displacement field \hat{v} resulting from these parameters is the optimal in terms of the best trade-off, according to the variational approach, between \mathcal{E}_{sim} and \mathcal{E}_{reg} . As addressed in [22], a scaling factor γ must be computed so that \mathcal{E}_{sim} and $\gamma \mathcal{E}_{reg}$ are comparable, since these functions do not have the same scale, the first being related to the intensities of the images and

the second to the smoothness of the deformation field. The scaling factor is given by the following expression:

$$\gamma = \left| \frac{\mathcal{E}_{sim,0} - \mathcal{E}_{sim,\infty}}{\mathcal{E}_{reg,0} - \mathcal{E}_{reg,\infty}} \right|, \tag{11}$$

where $\mathcal{E}_{sim,0}$ and $\mathcal{E}_{reg,0}$ are respectively the similarity energy and the regularization energy without any regularization (i.e. $\lambda = 0$) and $\mathcal{E}_{sim,\infty}$ and $\mathcal{E}_{reg,\infty}$ are the values of these energies for a large enough regularization parameter λ which makes the registration not appreciable (typically, $\lambda > 500$). Probably, the heuristic choice of the number of iterations \hat{n} is not optimal: it is almost certainly too small (i.e., the algorithm has not converged yet) or too large (i.e., the optimal registration could be reached sooner).

2. *Optimal parameters computation.* Our experiments over different types of images show that if the computed value of the quotient η between \hat{n} and $\hat{\lambda}$ is kept constant, the energies of the joint functional (1) show the same behavior as in figure Fig.1(a) for a large enough value of the regularization parameter (typically, $\lambda > 10$). To show the validity of the previous reasoning, figures Fig.1(b) and Fig.1(c) are presented. In the figure Fig.1(b) the number of iterations is four times higher than in figure Fig.1(a) (in this example, $\eta = \frac{100}{30} = \frac{400}{120} = 3.333$). The figure Fig.1(c) illustrates that η remains constant for every pair $(\hat{\lambda}, \hat{n})$ which minimizes (1). The idea is then to find a proportionality constant κ_o that allows for an optimal registration from the variational point of view and minimizes the number of iterations of the algorithm.

$$\lambda_o = \kappa_o \hat{\lambda}, \tag{12}$$

$$n_o = \kappa_o \hat{n} = \kappa_o \eta \hat{\lambda} = \eta \lambda_o. \tag{13}$$

This parameter κ_o is obtained as the minimum value of the proportionality parameter κ from which the slope of the joint functional (1) is close to zero ($< 10^{-6}$), i.e., the convergence has been reached, see figure Fig.1(d) (in the example, $\kappa_o = 5$). Thus, the optimal parameters λ_o and n_o can be calculated by employing equations (12) and (13).

4 Results

This section presents the results obtained with the presented methodology. Firstly, the method is validated in a difficult registration scenario, where the registration algorithm is very sensitive to the values of the simulation parameters. Then, the registration is performed on a medical image (obtained from [23]) under a synthetic deformation. Finally, the registration process is carried out for a pair of medical images, and once again the results are fully satisfactory.

In the first case, the registration is performed on a pair of photographic images in which an object was (physically) non-linearly deformed (figures Fig.2(a) and Fig.2(b)). The quotient calculated in the first step of the presented method is

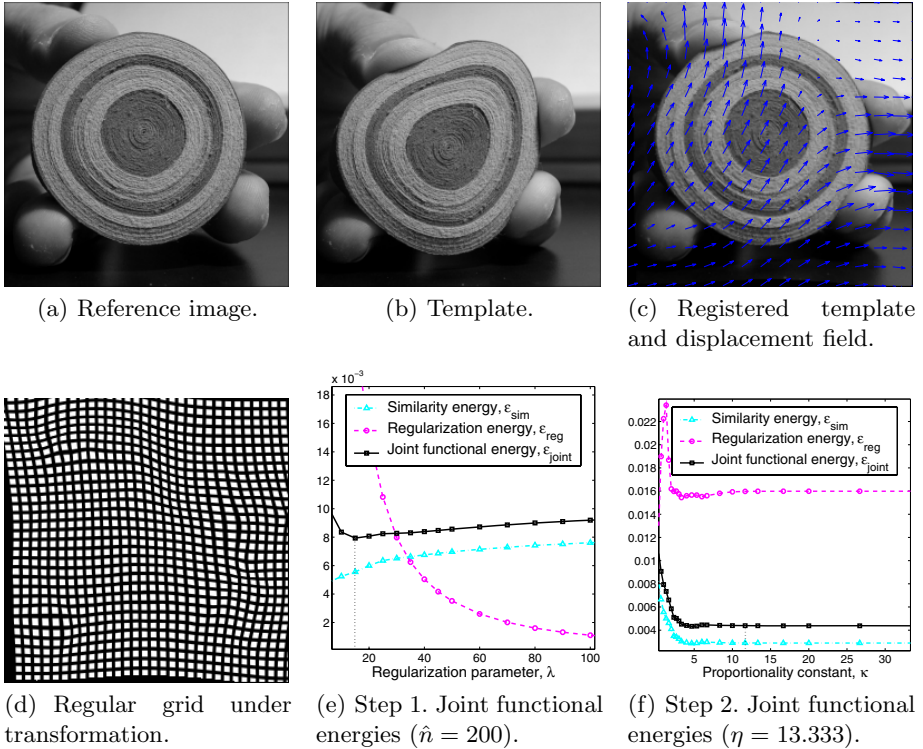


Fig. 2. Results for a real object with non-linear deformation

Table 1. Summary of the parameters involved in the first simulation (Fig.2)

Registration scheme	<i>Diffusion</i>
Number of iterations, \hat{n} (step 1)	200
Regularization parameter, $\hat{\lambda}$ (step 1)	15
Computed quotient, η (step 1)	13.333
Scaling factor, γ	0.048
Proportionality parameter, κ_o (step 2)	11.667
Optimal number of iterations, n_o (step 2)	2335
Optimal regularization parameter, λ_o (step 2)	175
PSNR before the registration	16.42 dB
PSNR after the registration	22.37 dB

$\eta = 13.333$ (corresponding to values of $\hat{n} = 200$ and $\hat{\lambda} = 15$, see figure Fig.2(e)). The computed scaling factor is in this case $\gamma = 0.048$. In the second step, we obtain the value of the proportionality constant $\kappa_o = 11.667$ (figure Fig.2(f)) so the optimal simulation parameters can be calculated using equations (12) and

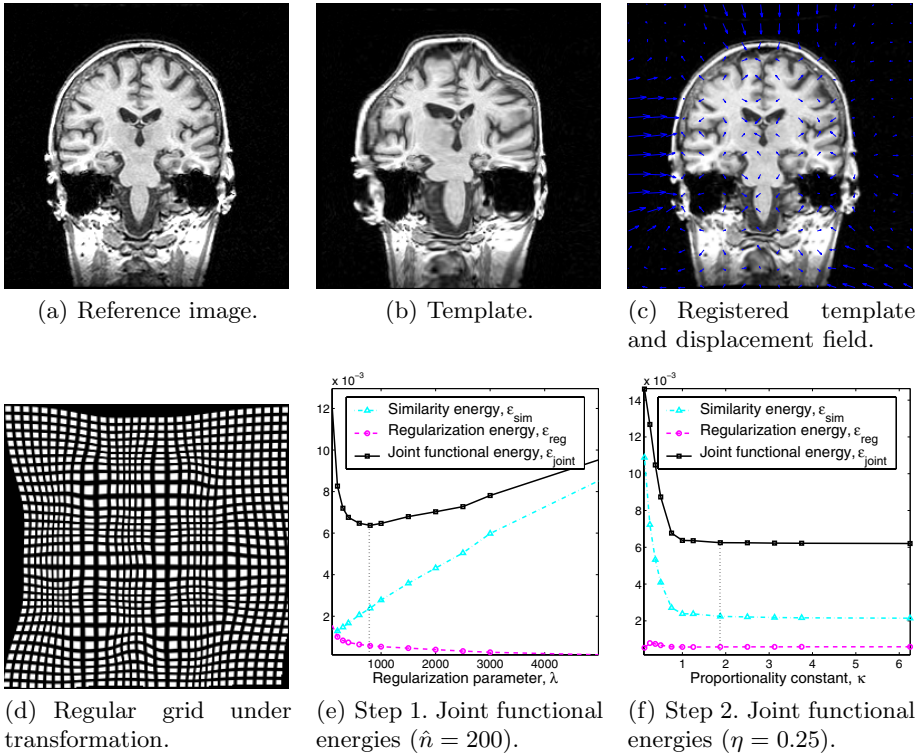


Fig. 3. Results for a medical image under synthetic deformation

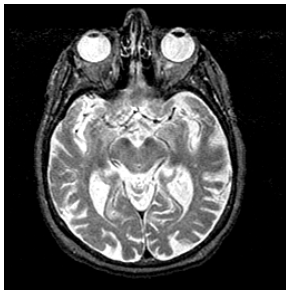
(13): $\lambda_o = 175$ and $n_o = 2235$. With these parameters, the registered image and the deformation field are obtained (see figure Fig.2(c)). On one hand the PSNR (peak signal-to-noise ratio) between the reference and registered images is quite high (22.37 dB versus 16.42 dB before registering), and on the other hand it is shown that a regular grid under the transformation $\mathbf{x} - \mathbf{v}(\mathbf{x})$ results in a visually smooth mesh (figure Fig.2(d)), so the trade-off between similarity and regularization energies can be fully appreciated. Table 1 summarizes the whole experiment.

In the second simulation, the aim is to register a magnetic resonance image (MRI) of a human brain, subject to a synthetic deformation field (figures Fig.3(a) and Fig.3(b)). The application of the proposed methodology is summarized in Table 2 and illustrated in figures Fig.3(e) and Fig.3(f). The optimally registered image and the matching vectors \mathbf{v} can be seen in figure Fig.3(c). The smoothness of a regular grid after applying these vectors can be appreciated in figure Fig.3(d).

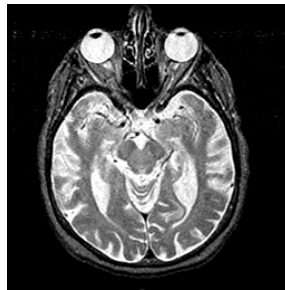
In the final simulation, the objective is to register two consecutive slices obtained from a MRI of a human brain (figures Fig.4(a) and Fig.4(b)). The two steps of the presented method can be seen in figures Fig.4(e) and Fig.4(f). The

Table 2. Summary of the parameters involved in the second simulation (Fig.3)

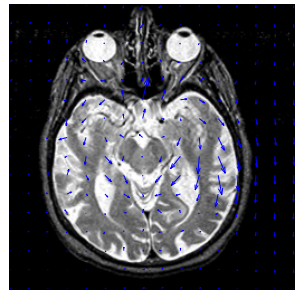
Registration scheme	Curvature
Number of iterations, \hat{n} (step 1)	200
Regularization parameter, $\hat{\lambda}$ (step 1)	800
Computed quotient, η (step 1)	0.25
Scaling factor, γ	6.883
Proportionality parameter, κ_o (step 2)	1.875
Optimal number of iterations, n_o (step 2)	375
Optimal regularization parameter, λ_o (step 2)	1500
PSNR before the registration	13.09 dB
PSNR after the registration	23.48 dB



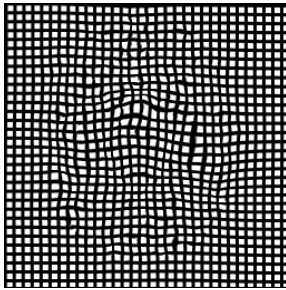
(a) Reference image.



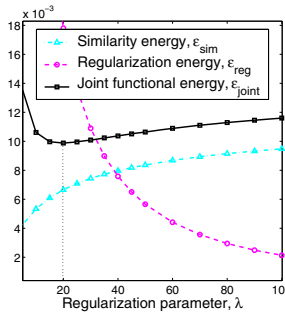
(b) Template.



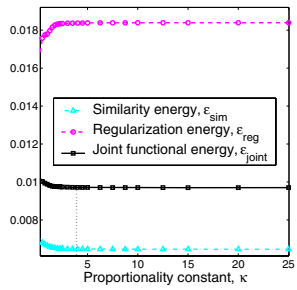
(c) Registered template and displacement field.



(d) Regular grid under transformation.



(e) Step 1. Joint functional energies ($\hat{n} = 200$).



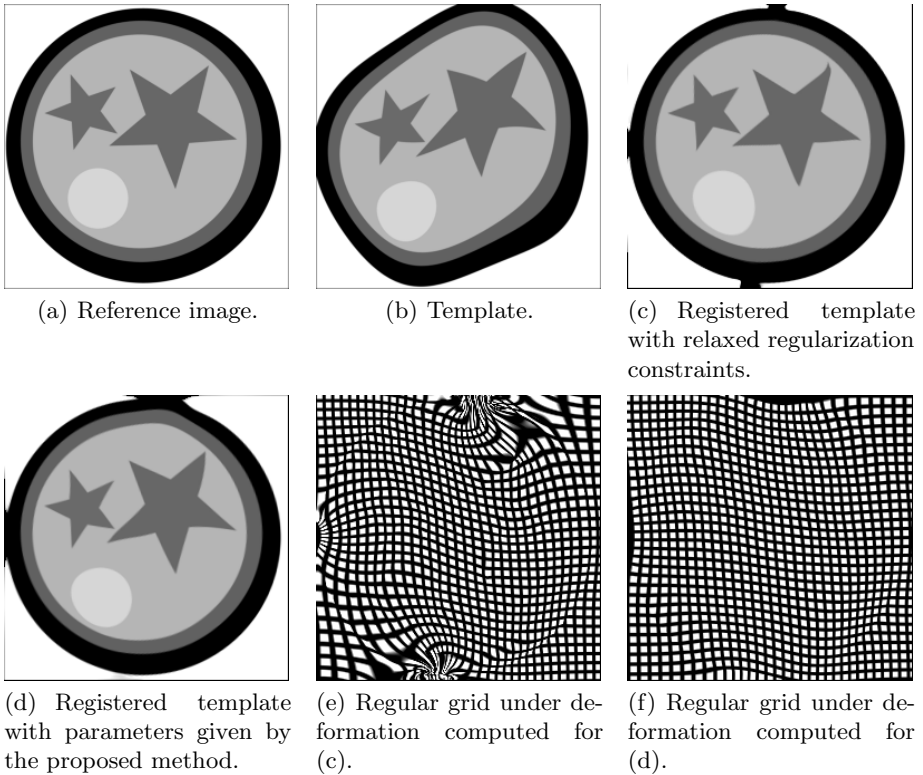
(f) Step 2. Joint functional energies ($\eta = 10$).

Fig. 4. Results for a pair of consecutive slices of a MRI scan

resulting registered image and the displacement field are shown in figure Fig.4(c). Figure Fig.4(d) shows that a regular grid under the computed transformation results in a smooth mesh. Table 3 summarizes this experiment.

Table 3. Summary of the parameters involved in the third simulation (Fig.4)

Registration scheme	<i>Diffusion</i>
Number of iterations, \hat{n} (step 1)	200
Regularization parameter, $\hat{\lambda}$ (step 1)	20
Computed quotient, η (step 1)	10
Scaling factor, γ	0.089
Proportionality parameter, κ_o (step 2)	4
Optimal number of iterations, n_o (step 2)	800
Optimal regularization parameter, λ_o (step 2)	80
PSNR before the registration	15.78 dB
PSNR after the registration	18.88 dB

**Fig. 5.** Registration with relaxed regularization constraints versus registration with optimal parameters

5 Discussion

In this paper, a two-step procedure is proposed to obtain, for non-parametric registration schemes, the optimal parameters and the minimum number of itera-

tions that achieve the best trade-off between similarity of the images and smoothness of the transformation. It is important to remark that the final transformation is always assumed to be a homeomorphism, see section 2. With this assumption the proposed values could be considered optimum in an *objective* sense. However, if the former assumption is not fulfilled it is possible to obtain best *subjective* results, in terms of visual quality (e.g. higher PSNR) of the registered image, in a sensitively lower number of iterations, and probably with different registration parameters (e.g. lower λ). To clarify this point, figure Fig.5 is shown (here, curvature registrations is performed). Note that in every registration method the final registration can always be *visually* improved (in this example, the PSNR is 2 dB higher) by relaxing the regularization constraints, but at the expense of a loss of smoothness and/or continuity in the computed mapping (see figure Fig.5(e) versus figure Fig.5(f)). In summary, the main goal of this work is to offer an objective upper limit of registration quality (with the smoothness and continuity requirements) and to provide the basic design guidelines to reach it.

References

1. Brown, L. G.: A survey of image registration techniques. *ACM Computing Surveys* **24**(4), (1992) 325–376.
2. Zitová, B. and Flusser, J.: Image registration methods: a survey. *Image and Vision Computing* **21**, (2003) 997–1000.
3. Maintz, J. and Viergever, M.: A survey of medical image registration. *Medical Image Analysis* **2**(1), (1998) 1–36.
4. Lester, H. and Arridge, S.: A survey of hierarchical non-linear medical image registration. *Pattern Recognition* **32**, (1999) 129–149.
5. Hajnal, J., Hill, D., and Hawkes, D.: *Medical image registration*. CRC Press, Boca Raton, FL (2001).
6. Goshtasby, A.: Registration of images with geometric distortions. *IEEE Transactions on Geoscience and Remote Sensing* **26**, (1988) 60–64.
7. Rohr, K.: *Landmark-based image analysis: using geometric and intensity models*. Computational Imaging and Vision Series, Kluwer Academic Publishers, Dordrecht **21** (2001).
8. Bajcsy, R. and Kovacic, S.: Multiresolution elastic matching. *Computer Vision, Graphics, and Image Processing* **46**(1), (1989) 1–21.
9. Bro-Nielsen, M. and Gramkow, C.: Fast fluid registration of medical images. *Lecture Notes in Computer Science* **1131**, (1996) 267–276.
10. Fischer, B. and Modersitzki, J.: Fast diffusion registration. M.Z. Nashed, O. Scherzer (eds), *Contemporary Mathematics 313, Inverse Problems, Image Analysis, and Medical Imaging*, AMS, (2002) 117–129.
11. Fischer, B. and Modersitzki, J.: Curvature based image registration. *Journal of Mathematical Imaging and Vision* **18**(1), (2003) 81–85.
12. Golub, G., Heath, M., and Wahba, G.: Generalized cross-validation as a method for choosing a good ridge parameter. *Technometrics* **21**, (1979) 215–223.
13. Fischer, B. and Modersitzki, J.: Large scale problems arising from image registration. *GAMM Mitteilungen* **27**(2), (2004) 104–120.

14. Ue, H., Haneishi, H., Iwanaga, H., and Suga, K.: Nonlinear motion correction of respiratory-gated lung SPECT images. *IEEE Transactions of Medical Imaging* **25**(4), (2006) 486–495.
15. Amit, Y.: A nonlinear variational problem for image matching. *SIAM Journal of Scientific Computing* **15**(1), (1994) 207–224.
16. Fischer, B. and Modersitzki, J.: Fast image registration - a variational approach. *Proceedings of the International Conference on Numerical Analysis & Computational Mathematics*, G. Psihoyios (ed.), Wiley, (2003) 69–74.
17. Zhang, Z., Jiang, Y., and Tsui, H.: Consistent multi-modal non-rigid registration based on a variational approach. *Pattern Recognition Letters* **27**, (2006) 715–725.
18. Horn, B. and Schunck, B.: Determining optical flow. *Artificial Intelligence* **17**, (1981) 185–204.
19. Thirion, J.-P.: Image matching as a diffusion process: an analogy with maxwell's demons. *Medical Image Analysis* **2**(3), (1998) 243–260.
20. Fischer, B. and Modersitzki, J.: A unified approach to fast image registration and a new curvature based registration technique. *Linear Algebra and its Applications* **308**, (2004) 107–124.
21. Braumann, U.-D. and Kuska, J.-P.: Influence of the boundary conditions on the results of non-linear image registration. *IEEE International Conference on Image Processing* **I**, (2005) 1129–1132.
22. Noblet, V., Heinrich, C., Heitz, F., and Armspach, J.-P.: Retrospective evaluation of a topology preserving non-rigid registration method. *Medical Image Analysis*, (2006) **In press**.
23. Johnson, K. and Becker, J.: The Whole Brain Atlas (1995-1999) (www.med.harvard.edu/aanlib/home.html).

Camera Calibration from a Single Frame of Planar Pattern

Jianhua Wang, Fanhuai Shi, Jing Zhang, and Yuncai Liu

Institute of Image Processing and Pattern Recognition,
Shanghai Jiao Tong University,
800 Dong Chuan Road, Shanghai 200240, P.R. China
{jian-hua.wang, fhshi, zhjseraph, whomliu}@sjtu.edu.cn

Abstract. A method to calibrate camera from a single frame of planar pattern is presented in this paper. For a camera model with four intrinsic parameters and visible lens distortion, the principal point and distortion coefficients are firstly determined through analysis of the distortion in an image. The distortion is then removed. Finally, the other intrinsic and extrinsic parameters of the camera are obtained through direct linear transform followed by bundle adjustment. Theoretically, the method makes it possible to analyze the calibration result at the level of a single frame. Practically, such a method provides a easy way to calibrate a camera used in industrial vision system on line and used in desktop vision system. Experimental results of both simulated data and real images validate the method.

1 Introduction

Camera calibration is a necessary step in 3D computer vision to extract metric information from 2D images, and therefore it has always been an important issue in photogrammetry[1,2,9,18] and computer vision[3,5,14,15,17,20]. Up to now, a number of methods have been developed to accommodate various applications. These methods can be classified into traditional calibration[3,4,5,7,8,11-17,19,20] and self-calibration[6,10].

Self-calibration is based on correspondences between image points from different views of a static scene, and no calibration objects are necessary. However, a large number of parameters need to be estimated, which results in complicated mathematical problem and hinders it from applicable use. In traditional calibration method, the dimension of the calibration object can be three[3,5], two[15] and one[19]. The three dimensional objects were used at beginning, but it is replaced by two dimensional objects because two dimensional objects are more convenient in making and using. Using one dimensional calibration objects can solve the occlusion problem in multi-camera system, but it requires several motions of the one dimensional calibration objects, so it is not convenient in practice. So far the popular methods use two dimensional calibration objects, i.e. planar pattern, in which several frames of the calibration objects at different poses and positions

are usually needed to calibrate a camera. On the one hand, it is desirable to calibrate vision system used in industry on line from just a single view of a planar pattern, on the other hand, the calibration of widely used desktop vision system also needs simple method, especially one from a single view of a planar pattern.

Besides above practical demand, camera calibration from a single frame of planar pattern has also theoretical significance. It provides a way to analyze the calibration result at the level of a single frame. The existing methods use multiple frames, and therefore the calibration result is from the cooperation of all the frames. If there are a few of outliers in the frames, their effects on the result are not easy to be found. On the contrary, if calibration can be done from just a single frame, the effect of an outlier on the calibration result can be easily found. If all the results from the calibration images are consistent with one another, it is sure that the calibration is good. In the sense, the method also provides a way to evaluate the calibration result.

Calibration methods using a single view has been mentioned in Tsai's and Zhang's methods[3,12,15], but all of them are based on the assumption that the principal point is at the image center, i.e. the frame center, which is different from the principal point in fact. Our method breaks through the assumption, and therefore goes forward a further step than Tsai's and Zhang's calibration methods using a single view.

In section 2, Our method is described in detail, including the determination of the principal point and lens distortion coefficients, determination of other intrinsic and extrinsic parameters, and a summary of the algorithm. In section 3, experiments on both computer simulated data and real images are described. Finally, in section 4 some conclusions and discussions are given.

2 Description of the Method

2.1 Main Idea

To calibrate a camera model with five intrinsic parameters including principal point (u_0, v_0) , principal length (α, β) and skew factor γ , at least three views of a planar pattern at different poses and positions are necessary[15]. For predominant majority of cameras used in vision system, the skew factor γ is small enough to be neglect due to the developed manufacturing technology of CCD device. So camera model with four parameters are commonly used, which needs two views of the planar pattern for calibration[15]. To our knowledge, this is the conclusion on minimum number of frames for calibration images reported in the literature.

The above conclusion are derived only from property of perspective projection. Is there any other information in an image which can help us to determine the parameters of a camera? Although the CCD device in a camera is good enough, lenses usually have visible distortion, especially the radial component. Usually, lens distortion is not desired in computer vision and the skilled in the art always manage to remove it. Can we use the lens distortion to help calibration?

Through theoretical derivation and experimental tests, we find that lens distortion can be used to simplify the camera calibration, and as a result the camera

calibration can be done from only a single frame of a planar pattern. Through analysis of the distortion in an image, the center and coefficients of lens distortion can be found, and the image can be corrected. The center of perspective projection on the retina, i.e. the principal point in camera model, is usually coincident with the center of lens distortion[8]. For a camera model with four parameters, after the principal point is known and the distortion is removed, the principal length can be obtained through direct linear transform[14,17]. Finally, the result can be fined through bundle adjustment.

2.2 The Camera Model

In order to describe our method clearly, we start with showing how a point in the view field is imaged onto the retina of a camera. The imaging process from the point P_w in the world coordinate system to the point P_d on the retina of a camera as shown in Figure 1 can be divided into four steps.

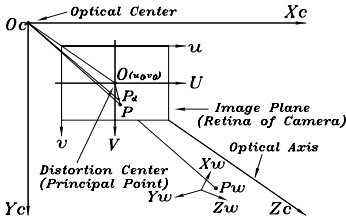


Fig. 1. The model of imaging

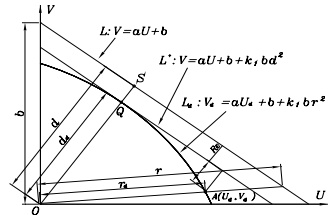


Fig. 2. Distortion of straight line

1) The world coordinates (X_w, Y_w, Z_w) of the point P_w is transformed to the camera coordinates $P_c = (X_c, Y_c, Z_c)$, which can be expressed as:

$$\begin{pmatrix} X_c \\ Y_c \\ Z_c \end{pmatrix} = R \begin{pmatrix} X_w \\ Y_w \\ Z_w \end{pmatrix} + T \tag{1}$$

where R is rotational matrix, and T is translational vector. Rotational matrix R can also be expressed as a rotational vector N, and they are related by Rodrigues Equation. Because the planar pattern lies in the $X_w Y_w$ plane, $Z_w = 0$, and equation (1) can be reduced as:

$$\begin{pmatrix} X_c \\ Y_c \\ Z_c \end{pmatrix} = \begin{pmatrix} r_1 & r_2 & T \end{pmatrix} \begin{pmatrix} X_w \\ Y_w \\ 1 \end{pmatrix} \tag{2}$$

where $r_i (i = 1, 2, 3)$ is the i th column of R. Let

$$Rt = \begin{pmatrix} r_1 & r_2 & T \end{pmatrix} \tag{3}$$

then

$$\begin{pmatrix} X_c \\ Y_c \\ Z_c \end{pmatrix} = Rt \begin{pmatrix} X_w \\ Y_w \\ 1 \end{pmatrix} \tag{4}$$

2) The point $P_c = (X_c, Y_c, Z_c)$ is imaged onto $P = (U, V)$ on the retina of a camera according to pinhole model, which can be expressed as:

$$\begin{pmatrix} U \\ V \\ 1 \end{pmatrix} = \frac{1}{Z_c} \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} X_c \\ Y_c \\ Z_c \end{pmatrix} \tag{5}$$

where α and β are the scale factors of perspective projection in the U and V direction respectively, which are also referred as principal length. Let

$$A = \begin{pmatrix} \alpha & 0 & 0 \\ 0 & \beta & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{6}$$

then

$$\begin{pmatrix} U \\ V \\ 1 \end{pmatrix} = \frac{1}{Z_c} A \begin{pmatrix} X_c \\ Y_c \\ Z_c \end{pmatrix} \tag{7}$$

Substitute (4) into (7),we get

$$s \begin{pmatrix} U \\ V \\ 1 \end{pmatrix} = AR_t \begin{pmatrix} X_w \\ Y_w \\ 1 \end{pmatrix} = H \begin{pmatrix} X_w \\ Y_w \\ 1 \end{pmatrix} \tag{8}$$

where

$$H = AR_t \tag{9}$$

and s is a scale factor.

3) The image point $P = (U, V)$ shifts to $P_d = (U_d, V_d)$ due to lens distortion, which can be expressed as:

$$\begin{pmatrix} U_d \\ V_d \end{pmatrix} = \begin{pmatrix} U \\ V \end{pmatrix} + \begin{pmatrix} \delta_U \\ \delta_V \end{pmatrix} \tag{10}$$

where δ_U and δ_V are lens distortions. Generally, lens distortion includes three components: radial distortion, decentering distortion and thin prism distortion [5]. But for usual vision system, the first order radial distortion is enough. So the lens distortion can be simply expressed as:

$$\begin{pmatrix} \delta_U \\ \delta_V \end{pmatrix} = \begin{pmatrix} k_1 \cdot U \cdot r^2 \\ k_1 \cdot V \cdot r^2 \end{pmatrix} \tag{11}$$

where k_1 is coefficient of the first order radial distortion, and r is the distance from the point P to the distortion center, which can be expressed as:

$$r^2 = U^2 + V^2 \tag{12}$$

4) The coordinates (U_d, V_d) of the image point is transformed into (u_d, v_d) in the pixel coordinate system, which can be expressed as:

$$\begin{pmatrix} u_d \\ v_d \end{pmatrix} = \begin{pmatrix} U_d \\ V_d \end{pmatrix} + \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} \tag{13}$$

where (u_0, v_0) is the coordinates of the principal point.

2.3 Estimation of the Principal Point

Substitute (11) into (10), we get:

$$\begin{pmatrix} U_d \\ V_d \end{pmatrix} = (1 + k_1 \cdot r^2) \begin{pmatrix} U \\ V \end{pmatrix} \tag{14}$$

$$r_d^2 = r^2(1 + k_1 \cdot r^2)^2 \tag{15}$$

and

$$\begin{pmatrix} U \\ V \end{pmatrix} = \frac{1}{1 + k_1 \cdot r^2} \begin{pmatrix} U_d \\ V_d \end{pmatrix} \tag{16}$$

Suppose the straight line L in Figure 2 comes from a straight line in the view field, and is described as:

$$V = aU + b \tag{17}$$

where a and b are constants to be determined. Substitute (16) into (17), we can get:

$$V_d = aU_d + b + k_1br^2 \tag{18}$$

If a line \overline{OS} is drawn from the principal point O such that it is perpendicular to the line L and intersects the curved line L_d at point Q , the line L' passing through the point Q and parallel to the line L can be expressed as:

$$V_d = aU_d + b + k_1bd^2 \tag{19}$$

where d is the distance from the principal point O to the straight line L . Suppose the distance from a point $A(U_d, V_d)$ on the distorted line L_d to the line L' is Re , then

$$Re = k_1d(d^2 - r^2) \tag{20}$$

If there are m straight lines $L_i (i = 1, 2, \dots, m)$ coming from the view field with n points $A_{ij}(U_{ij}, V_{ij}) (j = 1, 2, \dots, n)$ on each, the residue Re_i of the curved line L_{di} is the sum of distance Re_{ij} , i.e.

$$Re_i = \sum_{j=1}^n Re_{ij} = k_1d_i \sum_{j=1}^n (d_i^2 - r_{ij}^2) \tag{21}$$

From (21) we can see that the residue of a distorted line is proportional to the distance from the distortion center to corresponding undistorted straight line. Especially, if an undistorted straight line passes through the distortion center, the corresponding distorted line will remain straight. Therefore, if we set some straight lines in the view field and take a picture, the fitted straight line by Least Square Method which is nearest to the principal point will have the minimum residue. Suppose there are two sets of parallel lines in the view field, and the lines in the first set are orthogonal to the lines in the second set. Take a picture such that the lines in the picture are approximately vertical or horizontal. Suppose there are l approximately vertical lines and m approximately horizontal lines. Among the approximately vertical lines, the fitted line with the minimum residue is l_0 , and among the approximately horizontal lines, the fitted line with the minimum residue is m_0 . Then the intersection (\hat{u}_0, \hat{v}_0) of straight lines l_0 and m_0 can give an estimation of the principal point (u_0, v_0) . However, since

d_i and r_{ij} in equation (21) are related to the undistorted straight line, which are unknown, we can't use them. Fortunately, distortion is generally small, we can substitute d_i and r_{ij} with distorted values d_{d_i} and $r_{d_{ij}}$.

2.4 Estimation of the Distortion Coefficient k_1

For all lines in a picture, summing up the equation (21), we get,

$$\sum_{i=1}^{l+m} Re_i = k_1 \sum_{i=1}^{l+m} d_i \sum_{j=1}^n (d_i^2 - r_{ij}^2) \tag{22}$$

from which we can get an estimate \hat{k}_1 of k_1

$$\hat{k}_1 = \frac{\sum_{i=1}^{l+m} \sum_{j=1}^n Re_{ij}}{\sum_{i=1}^{l+m} d_i \sum_{j=1}^n (d_i^2 - r_{ij}^2)} \tag{23}$$

2.5 Refining of the Estimated Values

After we obtain the initial estimates of principal point and distortion coefficient k_1 , we can refine them through optimizing process according to the principle: in perspective projecting process, a straight line remains a straight line if and only if there is no lens distortion[16].

Suppose the maximum distance between two fitted straight lines from the approximately vertical lines which are next to each other is d_u , and the maximum distance between two fitted straight lines from the approximately horizontal lines which are next to each other is d_v , then we can refine the estimates (\hat{u}_0, \hat{v}_0) and \hat{k}_1 by minimizing the total residue expressed by formula (22) within the window from $(\hat{u}_0 - d_u, \hat{v}_0 - d_v)$ to $(\hat{u}_0 + d_u, \hat{v}_0 + d_v)$.

2.6 Determination of Other Parameters of the Camera

After the principal point (u_0, v_0) and distortion coefficient k_1 are determined, we can obtain the principal length α, β , rotation matrix R , and translational vector T as follows. From equation (13) we get

$$\begin{pmatrix} U_d \\ V_d \end{pmatrix} = \begin{pmatrix} u_d \\ v_d \end{pmatrix} - \begin{pmatrix} u_0 \\ v_0 \end{pmatrix} \tag{24}$$

From equation (24), (15) and (16) we can remove distortion and get the undistorted image coordinates (U, V) of the point P . From the coordinates (U, V) of 4 intersections formed by two approximately vertical lines and two approximately horizontal lines in the image system and the coordinates (X_w, Y_w) of their corresponding points in the world system, the homography H in equation (8) can be obtained up to a scale factor λ , i.e.

$$\lambda \begin{pmatrix} h_1 & h_2 & h_3 \end{pmatrix} = A \begin{pmatrix} r_1 & r_2 & T \end{pmatrix} \tag{25}$$

where h_1, h_2, h_3 is the column vector of H . We can arrange many approximately vertical lines and approximately horizontal lines and get a lot of intersections,

and therefore obtain the homography H through Linear Least Square Method. Once the homography H is obtained, we can get

$$r_1 = \lambda A^{-1}h_1 \tag{26}$$

$$r_2 = \lambda A^{-1}h_2 \tag{27}$$

Because R is an orthonormal matrix, we have

$$h_1^T A^{-T} A^{-1} h_2 = 0 \tag{28}$$

$$\lambda^2 h_1^T A^{-T} A^{-1} h_1 = 1 \tag{29}$$

$$\lambda^2 h_2^T A^{-T} A^{-1} h_2 = 1 \tag{30}$$

From equation (6), we get

$$A^{-1} = \begin{pmatrix} \frac{1}{\alpha} & 0 & 0 \\ 0 & \frac{1}{\beta} & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{31}$$

$$A^{-T} A^{-1} = \begin{pmatrix} \frac{1}{\alpha^2} & 0 & 0 \\ 0 & \frac{1}{\beta^2} & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{32}$$

In equation (28),(29) and (30) there are 3 unknown α, β and λ , they can be found from H. After α, β and λ are obtained, we can compute r_1 and r_2 from equation (26) and (27), then compute r_3 and T from the following equations

$$r_3 = r_1 \times r_2 \tag{33}$$

$$T = \lambda A^{-1}h_3 \tag{34}$$

So far all the intrinsic and extrinsic parameters of a camera have initial values, and we can refine them by bundle adjustment.

2.7 Summary of the Algorithm

The method described above can be summarized as the following algorithm:

- 1) Arrange two sets of parallel lines in the view field, such that the lines in the first set are orthogonal to the lines in the second set, such as a checkerboard;
- 2) Take a picture such that the lines in the picture are approximately vertical or horizontal. Suppose l lines are approximately vertical and m lines are approximately horizontal;
- 3) Fitted out a straight line from each of the lines in a picture with the Least Square Method;
- 4) Suppose the line l_0 has the minimum residue among the approximately vertical lines, and the line m_0 has the minimum residue among the approximately horizontal lines. Then the intersection (\hat{u}_0, \hat{v}_0) of straight lines l_0 and m_0 can give an estimation to the principal point (u_0, v_0) .
- 5) Estimate the distortion coefficient k_1 with formula (23);
- 6) Refining the estimates of the principal point and the distortion coefficient;
- 7) Find the principal length α, β , rotation matrix R , and translational vector T using direct linear transform method;
- 8) Refine all parameters by bundle adjustment.

It should be noted that the part of estimating the distortion center in the algorithm including step 1) to 4) is similar to the steps i) to iii) in [21], which focused on modelling variable resolution imaging system. But in this paper, the method is supported by mathematic derivation in detail.

3 Experiment Result

3.1 Test on Computer Simulated Data

For camera calibration, the true parameters are unknown. To verify our method, we predetermine a set of intrinsic and extrinsic parameters of an imagined camera, from which an image of a checkerboard with 18×26 blocks in the view field is generated. The size of the blocks is $14mm \times 14mm$. Noise normally distributed with 0 mean and σ standard deviation is added to the corner points. Then we calibrate the imagined camera with the corner points in the checkerboard image using our method and compare the results with the given values.

Table 1. The given intrinsic parameters and calibrated results at noise level 0.6 pixels, including mean and standard deviation of 49 calibrated values

Parameter	u_0	v_0	α	β	k_1
Given	400	260	960	960	-0.5
Mean	400.08	260.11	959.76	960.34	-0.4999
Std.	0.84	0.55	32.12	34.83	0.0024

The intrinsic parameters are given in Table 1, which are typical for cameras with 8mm lens. To examine the effect of different pose and position of the checkerboard upon the calibration accuracy, we predetermine 49 cases of extrinsic parameters, as shown in Figure 3. We divide the 49 cases into seven groups. The first four groups are designed for examining the effect of rotation vector N upon calibration accuracy, and the last three groups are for translational vector T . In group one including case 1 to 7, we increase the first component n_1 of the rotational vector N from $10 \times 180/\pi$ to $40 \times 180/\pi$ with an increment of $5 \times 180/\pi$, while the other parameters are fixed. In group two including case 8 to 14, we increase the second component n_2 of the rotational vector N from $10 \times 180/\pi$ to $40 \times 180/\pi$ with an increment of $5 \times 180/\pi$, while the other parameters are fixed. In group three including case 15 to 21, we increase the first component n_1 and the second component n_2 of the rotational vector N simultaneously from $10 \times 180/\pi$ to $40 \times 180/\pi$ with an increment of $5 \times 180/\pi$, while the other parameters are fixed. In group four including case 21 to 28, we increase the first component n_1 of the rotational vector N from $10 \times 180/\pi$ to $40 \times 180/\pi$, at the same time decrease the second component n_2 of the rotational vector N from $40 \times 180/\pi$ to $10 \times 180/\pi$, both with the same increment of $5 \times 180/\pi$, while the other parameters are fixed. In group five including case 29 to 35, we increase the first component t_1 of the translational vector T from $-300mm$ to $-180mm$ with an increment of $20mm$, while the other parameters are fixed. In group six

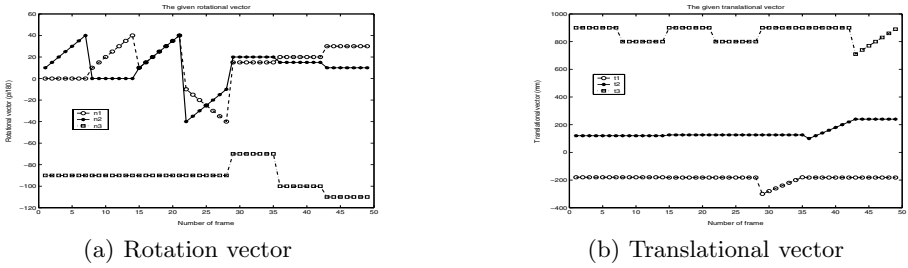


Fig. 3. The given extrinsic parameters

including case 36 to 42, we increase the second component t_2 of the translational vector T from $100mm$ to $220mm$ with an increment of $20mm$, while the other parameters are fixed. In group seven including case 42 to 49, we increase the third component t_3 of the translational vector T from $710mm$ to $890mm$ with an increment of $30mm$, while the other parameters are fixed.

We did calibration with the simulated data at noise level of $\sigma = 0.2, 0.4, 0.6, 0.8$ and 1.0 pixels respectively. Figure 4 shows the calibration results with noise level $\sigma = 0.6$ pixel, with the mean and standard deviation of the calibrated intrinsic parameters listed in Table 1 to be compared with the given values. The change of deviation with noise level is shown in Figure 4(f). When the noise level is zero, the deviation of all parameters in all 49 cases are zero, which proves that our method is theoretically correct. In general, the proposed method can correctly calibrate out the intrinsic and extrinsic parameters of the supposed camera. In group three from case 15 to case 21, the deviation of principal length α, β and the translational component t_3 are big, especial at case 15 and 16, which indicate that these poses are harmful to calibration accuracy. Except the case 15 and 16, the deviations of the other 47 calibrations are in a acceptable range when the noise level is within $\sigma = 1.0$ pixel, which indicates that the method can be used with real images, because the accuracy of extracting the corner points from a checkerboard image is within sub-pixel. As for the special poses, such as case 15 and 16, we can manage to avoid them in taking images for calibration. The above results encourage us to do calibration with real images.

3.2 Test on Real Image

To verify the proposed method, we calibrated two cameras with $8mm$ lens and $6mm$ lens respectively which are typically used. Both cameras are WAT 902B CCD camera with resolution of 768×576 and unit cell size of CCD sensor being $8.3\mu m \times 8.3\mu m$. Both lenses are standard CCTV lenses.

For each camera, 16 pictures of a checkerboard at different poses and positions were taken as shown in Figure 5(a) and Figure 6(a). The calibration results are shown in Figure 5 and Figure 6. To judge the calibration results, we calibrated the two cameras with Zhang’s method [15] and the results are listed in Table 2 together

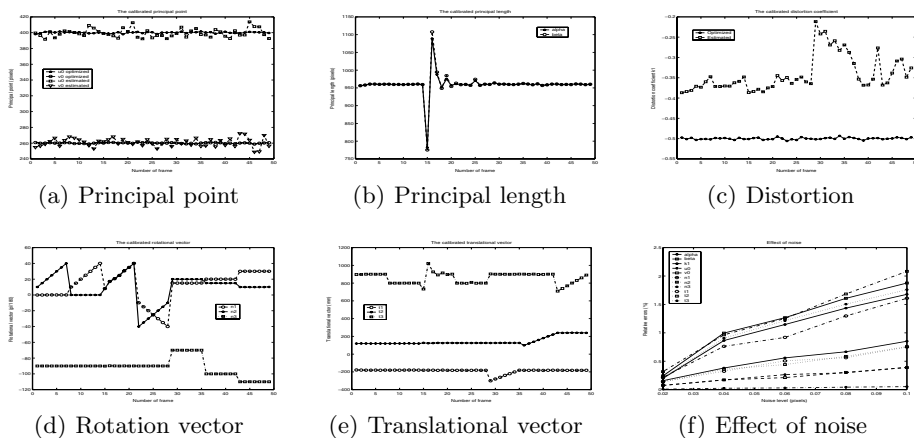


Fig. 4. The calibration results at noise level $\sigma = 0.6$ pixels

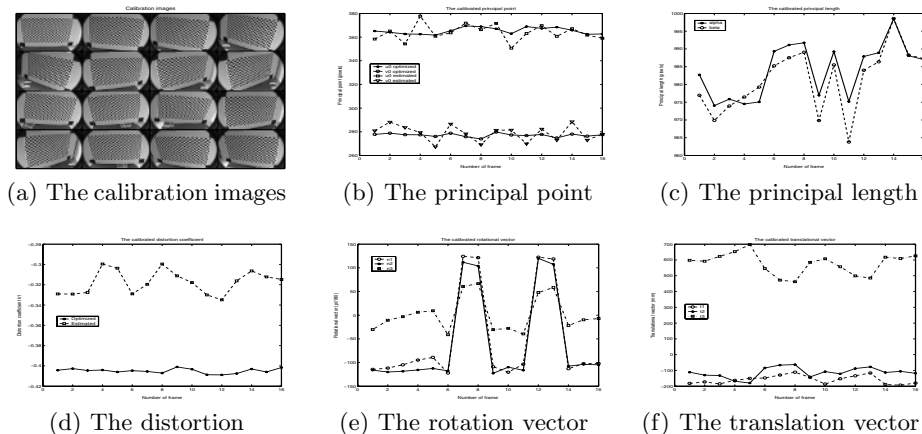


Fig. 5. The calibration results of camera with 8mm lens

Table 2. Intrinsic parameters of cameras calibrated from our method and Zhang's method

Camera	<i>with lens of 8mm focal length</i>					<i>with lens of 6mm focal length</i>				
Parameter	u_0	v_0	α	β	k_1	u_0	v_0	α	β	k_1
Mean	365.32	277.17	984.13	981.37	-0.4049	353.79	255.71	733.30	732.67	-0.3391
Std.	2.72	1.50	7.77	9.01	0.0023	2.51	2.03	10.68	12.81	0.0027
Zhang'	361.34	279.19	989.69	987.22	-0.4064	350.60	258.29	734.97	732.53	-0.3396

with the mean and standard deviation of the 16 calibrations from our method. Comparing the results in Figure 5, Figure 6 and in Table 2, we can see that:

1) the average of the 16 calibrations is close to the results from Zhang's method;

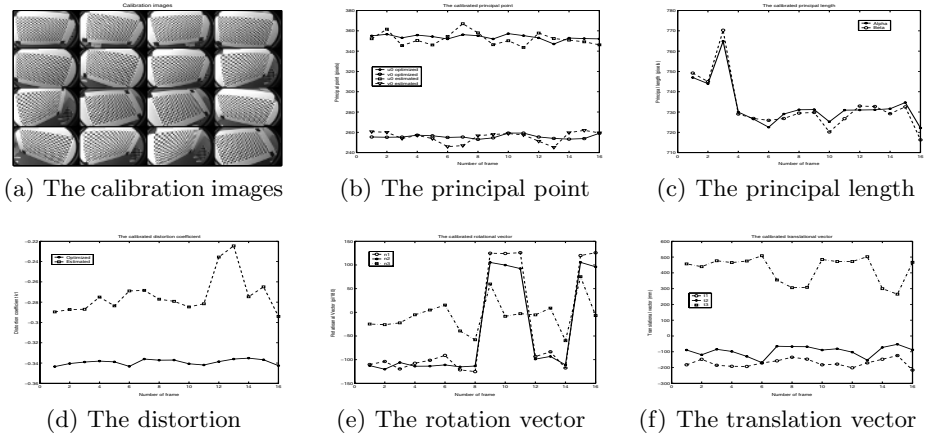


Fig. 6. The calibration results of camera with 6mm lens

2) for each calibration, the maximum relative errors of the calibrated parameters are all within 2%.

For usual vision system, the calibration accuracy from our method is acceptable. Because only one view of the checkerboard is needed, our method is simple and fast. If multiple images are taken, not only can the average be closer to the true value, but also the outlier among them can be easily found and removed.

4 Conclusion and Discussion

In this paper, we present a method to calibrate camera from a single frame of a planar pattern. For a camera model with four intrinsic parameters and visible lens distortion, the principal point and the distortion coefficient are firstly determined through analysis of the distortion in an image. Then the distortion can be removed. Finally, the other intrinsic and extrinsic parameters of the camera are obtained through direct linear transform followed by bundle adjustment. Experimental results of both simulated data and real images show that the calibration accuracy from our method is acceptable for usual vision system.

Theoretically, the method makes it possible to analyze the calibration result at the level of a single frame. Practically, such a method provides a easy way to calibrate a camera used in industrial vision system on line and used in desktop vision system. In this paper, the real images in experiment are not yet in real time. Now we are working on calibration using images in real time.

It should be noted that our method is based on the assumption that the principal point and the center of distortion are coincident[8]. In fact, a predominant majority of calibration methods so far, including Tsais[3] and Zhangs[15,19], are based on the assumption. Recently it has been pointed out that the principal point and the center of distortion are not coincident[20]. However, for usual vision system which does not require high accuracy, the difference between the principal point and the center of distortion can be neglected.

References

1. D.C. Brown. Close-Range Camera Calibration. *Photogrammetric Eng.*, vol. 37, no. 8, 1971.
2. W. Faig. Calibration of Close-Range Photogrammetry Systems: Mathematical Formulation. *Photogrammetric Eng and Remote Sensing*, vol. 41, no. 12, 1975.
3. R.Y. Tsai. A Versatile Camera Calibration Technique for High-Accuracy 3D Machine Vision Metrology Using Off-the-Shelf TV Cameras and Lenses. *IEEE J. Robotics and Automation*, vol. 3, no.4, pp. 323-344, Aug.1987.
4. B. Caprile and V. Torre. Using Vanishing Points for Camera Calibration. *Intl J. Computer Vision*, vol. 4, no. 2, pp. 127-140, Mar.1990.
5. J. Weng. P. Cohen, and M. Herniou. Camera Calibration with Distortion Models and Accuracy Evaluation. *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 14, no. 10, pp. 965-980, Oct. 1992.
6. O. Faugeras, T. Luong, and S. Maybank. Camera Self-Calibration: Theory and Experiments. *Proc Second European Conf. Computer Vision*, G. Sandini, ed., vol. 588, pp. 321-334, May 1992.
7. G.Q.Weï and S.D. Ma. A Complete Two-Plane Camera Calibration Method and Experimental Comparisons. *Proc. Fourth Intl Conf. Computer Vision*, pp. 439-446, May 1993.
8. R.G.Willson and S.A.Shafer. What is the center of the image ?. *Technical Report CMU-CS-93-122*, Carnegie Mellon University,1993.
9. T.A. Clarke and J.G. Fryer. The Development of Camera Calibration Methods and Models. *Photogrammetric Record*,16(91): 51-66, April 1998.
10. B. Triggs. Autocalibration from Planar Scenes. *Proc. Fifth European Conf. Computer Vision*, pp. 89-105, June 1998.
11. D. Liebowitz and A. Zisserman. Metric Rectification for Perspective Images of Planes. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 482-488, June 1998.
12. I. Shimizu, Z. Zhang, S. Akamatsu, and K. Deguchi. Head pose determination from one image using ageneric model.*Int'l Conf. on Automatic Face and Gesture Recognition*,1998.
13. P. Sturm and S. Maybank. On Plane-Based Camera Calibration: A General Algorithm,Singularities, Applications. *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, pp. 432-437, June 1999.
14. R. Hartley and A. Zisserman. Multiple View Geometry in Computer Vision. *Cambridge*, 2000.
15. Z. Zhang. A Flexible New Technique for Camera Calibration. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 22(11):1330-1334,2000.
16. Frederic Devernay,Olivier Faugeras. Straight lines have to be straight. *Machine Vision and Applications*, (2001) 13: 14-24.
17. A. Heyden and M. Pollefeys. *Multiple View Geometry, Emerging Topics in Computer Vision*, G. Medioni and S.B. Kang, eds., Prentice Hall, 2003.
18. Chris McGlone, Edward Mikhail and James Bethel. *Manual of Photogrammetry*, 5th Edition, 2004.
19. Zhengyou Zhang. Camera calibration with onedimensional objects. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 26(7):892 - 899, July. 2004.
20. R.I. Hartley and S.B. Kang. Parameter-free radial distortion correction with centre of distortion estimation. *ICCV*,2005.
21. Anup Basu and Sergio Licardio. Alternative models for fish-eye lenses. *Pattern Recognition Letters*,16(1995)433-441.

Stereo Matching Using Scanline Disparity Discontinuity Optimization

Ho Yub Jung, Kyoung Mu Lee, and Sang Uk Lee

School of Electrical Eng., ASRI, Seoul National University,
151-600, Seoul, Korea
hoyub@diehard.snu.ac.kr, kyoungmu@snu.ac.kr,
sanguk@sting.snu.ac.kr

Abstract. We propose a scanline energy minimization algorithm for stereo vision. The proposed algorithm differs from conventional energy minimization techniques in that it focuses on the relationship between local match cost solution and the energy minimization solution. The local solution is transformed into energy minimization solution through the optimization of the disparity discontinuity. In this paper, disparity discontinuities are targeted during the energy minimization instead of the disparities themselves. By eliminating and relocating the disparity discontinuities, the energy can be minimized in iterations of $O(n)$ where n is the number of pixels. Although dynamic programming has been adequate for both speed and performance in the scan-line stereo, the proposed algorithm was shown to have better performance with comparable speed.

1 Introduction

A dense disparity map is obtained by finding the correct corresponding pixels from two images of same scene. Various different techniques and their performances were evaluated by Scharstein and Szeliski [12]. According to their ongoing survey on new stereo schemes, the energy minimization solution is one of the common frameworks for current stereo vision techniques. And with the introduction of such techniques as graph cuts (GC) and belief propagation (BP), the disparity map of test images can be obtained with high accuracy [4] [13] [6].

However, the performance of the recent top ranking algorithms has become somewhat saturated. This is evident with introduction of new test images with more difficult features in the Middlebury website [1]. Although many researchers are still looking for that small margin of improvement that enable them to rank at top, others focused their attention on the faster algorithm with reasonable trade off in performance. Tree dynamic programming (DP) and reliability DP were introduced recently to CVPR and PAMI [14] [8]. Both of the techniques emphasized the faster speed with sufficient accuracies. And like the previous papers mentioned, in this paper, the contribution comes from faster speed and competitive performance from the proposed energy minimization technique.

1.1 Related Work

DP is most closely related to the proposed method for its scanline domain. DP finds the minimum path across disparity image space by first finding the minimum paths from the left, and tracing the shortest path back from other side [3] [7] [10]. DP is one of the older techniques and, other than the fast computational time, the performance is ranked among lowest.

However, recently DP has shown to be highly accurate if the matching cost can be aggregated properly. Kim *et al.* showed that by using rotational rod filter, a top ranking disparity map can be obtained, comparable to many of the global methods [11]. Although their method was fast by applying DP, the complicated match cost aggregation still hindered over all speed, going over few seconds.

Contrast, tree DP and reliability DP performs noticeably worse than [11], but they were able to keep the computation time to a fraction of second. Veckler *et al.* organized the image into a tree path and minimized the energy using DP. By enhancing the unambiguous match points, Gong *et al.* came up with reliability DP, that computes disparity map in semi-real time with help of add on hardware. And 5 state DP takes in account of the slanted surface also improved the original DP significantly awhile keeping the computational time small [1] [10]. Semi-global technique is also very closely related to DP, where the disparities are obtained according to the shortest multiple paths [9]. Although forementioned algorithms are additions to DP, they are most resembling to the proposed method in speed and performance

2 Disparity Discontinuity Optimization and Energy Function

The proposed algorithm can be described in two steps. First the initial disparity map is obtained without incorporating the discontinuity cost. Second, the disparity map is optimized iteratively according to the energy function. In this section, we will explain how the energy is minimized for each disparity discontinuity.

2.1 Energy Function

We start with the formulation of the energy function using the same notation from [4] and [6].

$$E(f) = \sum_{(p,q) \in N} V(f_p, f_q) + \sum_{p \in P} D_p(f_p). \quad (1)$$

As mentioned before, this is a scan line algorithm; therefore, the neighborhood N will consists only of adjacent pixels left and right. $V(f_p, f_q)$ is the disparity discontinuity between point p and q . $D(f_p)$ is the matching cost of labeling f on point p .

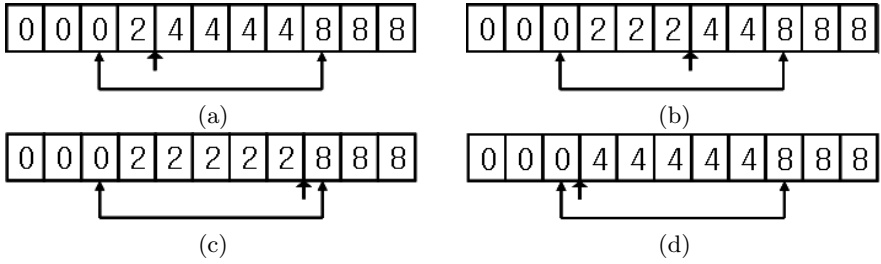


Fig. 1. An example of scanline disparity map, discontinuity, and local range during optimization process are shown. (a) shows the initial disparity map, and (b), (c), and (d) show next possible discontinuity position. (a) An initial disparities are indicated by integers. The targeted discontinuity and effected pixel ranges are shown with arrows. (b) A different disparity discontinuity position might yield smaller energy value. The disparity value changes as the discontinuity position changes. (c) Another possible discontinuity position. In this figure, the discontinuity is effectively eliminated. (d) Another discontinuity position that can eliminate the discontinuity. The new position might result in smaller energy value calculated within the boundary indicated by the arrows.

The complication for energy minimization comes from the discontinuity cost. Now, for simplicity, let us examine the data term in (1) only.

$$E_d(f) = \sum_{p \in P} D_p(f_p). \tag{2}$$

The minimum of the above equation is trivial, since the local minimum is also the global minimum. An apparent observation is that the minimization of equation (2) will have more disparity discontinuity than the minimum solution of equation (1). From this simple and obvious apprehension, we can make a following assertion; the minimum solution to equation (2) can approximately be transformed into the minimum solution of equation (1) by the elimination and relocation of disparity discontinuity. And for the 1-D image, finding the local optimal discontinuity position becomes a trivial task as will be shown in following sections. And for the clarification, from now on, the global minimization in this paper will be referring to the minimum energy function along a single scanline image, and not the whole image.

2.2 Disparity Discontinuity Optimization (DDO)

When we approach the stereo problem as the optimization of discontinuity instead of disparity, discontinuity, in all likelihood, will have fewer number than the disparities at each pixels. Nevertheless, the reduced range will not help the computational speed when the global minimization still remains NP hard. Therefore, the local disparity discontinuity optimization will be adopted, and in practice local optimization works very well.

Given a particular disparity discontinuity to optimize, the local boundary is defined by the two closest discontinuities in the left and right side of the discontinuity at question. A local boundary can be defined for a discontinuity between (p, q) , where p and q are adjacent pixels. Let A_p be the group of pixel position including pixel p having same disparity values and within discontinuity borders and same for A_q . For a discontinuity at (p, q) , the local boundary can be defined as $A_p \cup A_q$. And example is shown on Fig. 1. We also define r and s as the minimum and maximum 1D positions.

$$r \leq A_p \cup A_q \leq s. \quad (3)$$

The size of local boundary is dynamically changing from each discontinuity, but the locality size is inversely proportional to the number of the disparity discontinuities and so computational work remains steady across the scan line.

Finding the local optimal position/elimination of a discontinuity as mentioned before becomes a trivial task in 1D. Exact energy value can be calculated for each possible new discontinuity position within the local boundary.

Fig. 1 gives an general idea of the method. Initially, A_p will have same disparity values, which is numerically indicated as constant L . A_q will have disparity value R also constant. Let $D_x(L)$ represent the matching cost at pixel position x with disparity value L . $V(f_x, f_{x+1})$ will denote a discontinuity cost at points x and $x + 1$. If a discontinuity is found on pixels p^i and $p^i + 1$, the next discontinuity position between p^{i+1} and $p^{i+1} + 1$, with smaller energy cost, can be found with following equation.

$$p^{i+1} = \arg \min_{r+1 \leq k \leq s} \left(\sum_{x=r}^k D_x(L) + \sum_{x=k+1}^s D_x(R) + \sum_{x=r-1}^s V(f_x, f_{x+1}) \right). \quad (4)$$

The k with smallest energy cost is chosen as the next discontinuity position. But, the discontinuity is effectively eliminated when the, new position p^{i+1} becomes $r + 1$ or s . The equation (4) is simply the local minimum of the energy equation of (1).

2.3 Computational Work and Heuristics

For each discontinuity at p^i , the energy value at each position along $A_p \cup A_q$ are being calculated. And as mentioned before, the size of $A_p \cup A_q$ decreases as the number of discontinuity increases and vice versa. Thus the total number of next possible position on a scanline, regardless of the number of discontinuities, will be $2n$, where n is the number of pixels. The optimization or reposition of all the discontinuity in a scanline is counted as single iteration and such process was shown to be $O(n)$. The physical computational time is shown in Table 1.

Although it may be somewhat trivial, we will go over the sequential calculation of the equation (4) over the range $A_p \cup A_q$. $E(i)$ stands for the local energy function with discontinuity between pixel position i and $i + 1$. Except for special

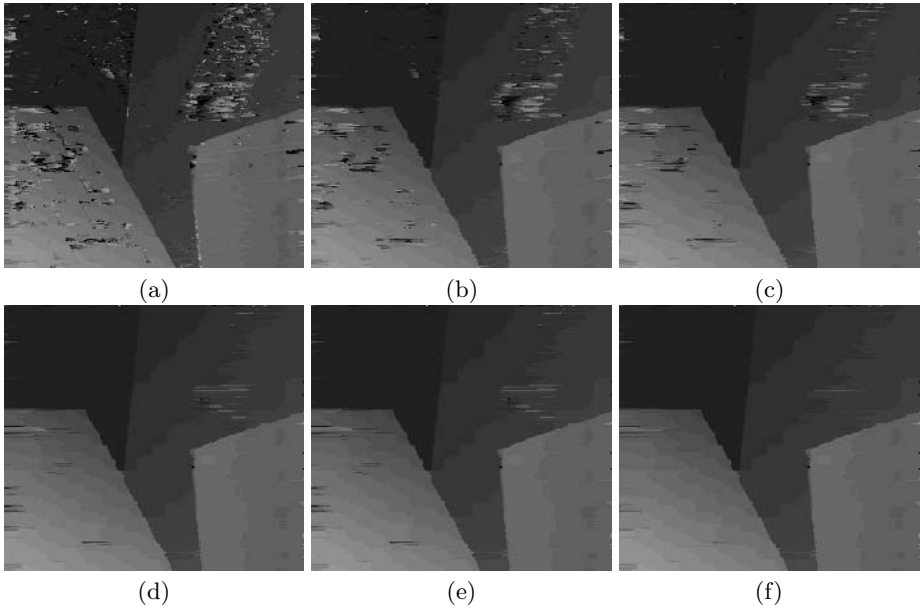


Fig. 2. The disparity discontinuity optimization on Venus image after (a) 0 iteration (b) 5 iteration (c) 10 iteration (d) 18 iteration (e) 25 iteration (f) 50 iteration

cases where i is equal to $r + 1$ or $s - 1$, we can easily see that $E(i + 1)$ can be obtain from $E(i)$ as shown in the equation below.

$$E(i + 1) = E(i) - D_{i+1}(L) + D_{i+1}(R). \quad (5)$$

In practice, however, the optimization was proceeded incrementally to prevent the possible early eliminations of the correct disparity. Given the old discontinuity position at p^i and the new position at p^{i+1} , obtained by equation (4), the incremental position can be written as

$$p_{incremental}^{i+1} = \arg \min_{k=p^i+1, p^i-1, p^i} |k - p^{i+1}|. \quad (6)$$

The new discontinuity position will be either $p^i + 1$, $p^i - 1$, or p^i , which is one/zero pixel toward the calculated minimum energy position from the old position. The incremental optimization is shown in Fig. 2.

Additionally, the initial matching cost for equation (2) was calculated differently from the matching aggregation for energy minimization equation (1). A tall window was used to degrade the streaking effect during energy minimization steps, but a wide window was used for initial disparity map to eliminate the vertical errors that may be present if the same tall window were to be used initially. The details of match cost aggregation will be presented in the next section.

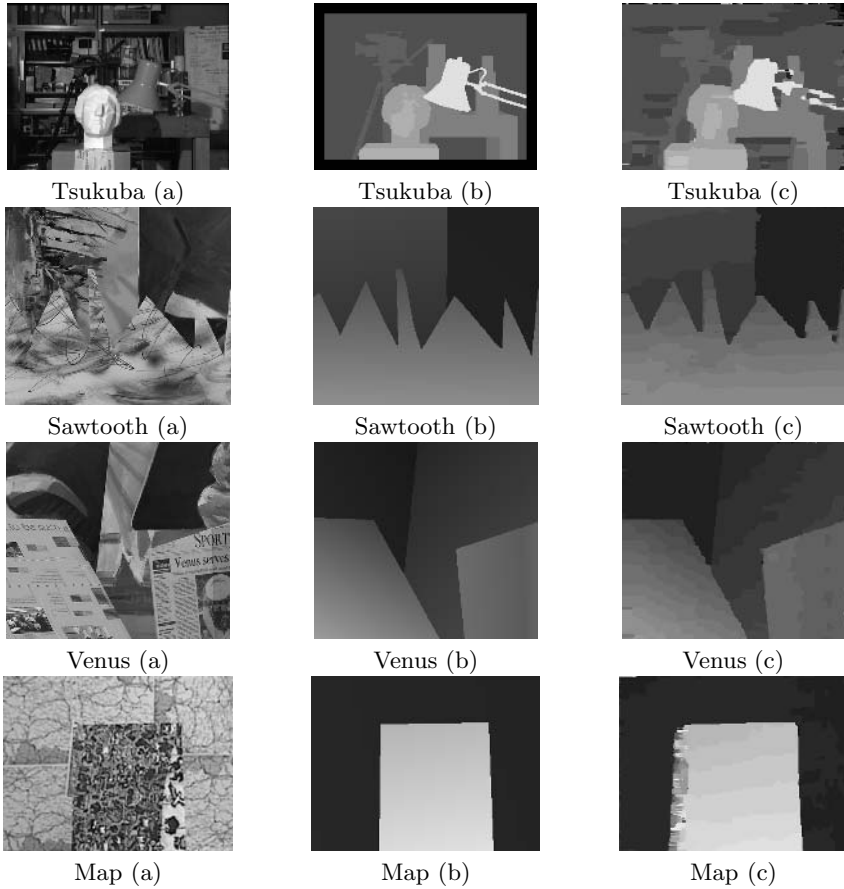


Fig. 3. Disparity maps produced by the proposed method are shown for Tsukuba, Sawtooth, Venus, and Map image. (a) Reference image (b) Ground truth (c) Disparity map after energy convergence.

3 Experiments

3.1 Parameters

The proposed algorithm was evaluated on 6 test images proposed by Scharstein and Szeliski, provided on the web [12] [1]. The percentages of wrong disparity with greater than the difference of 1 were used as the metric for evaluation.

The discontinuity cost between point p and q were found with following equation, taking $c = 2.0$.

$$V(f_p, f_q) = |f_p - f_q| \cdot c. \quad (7)$$

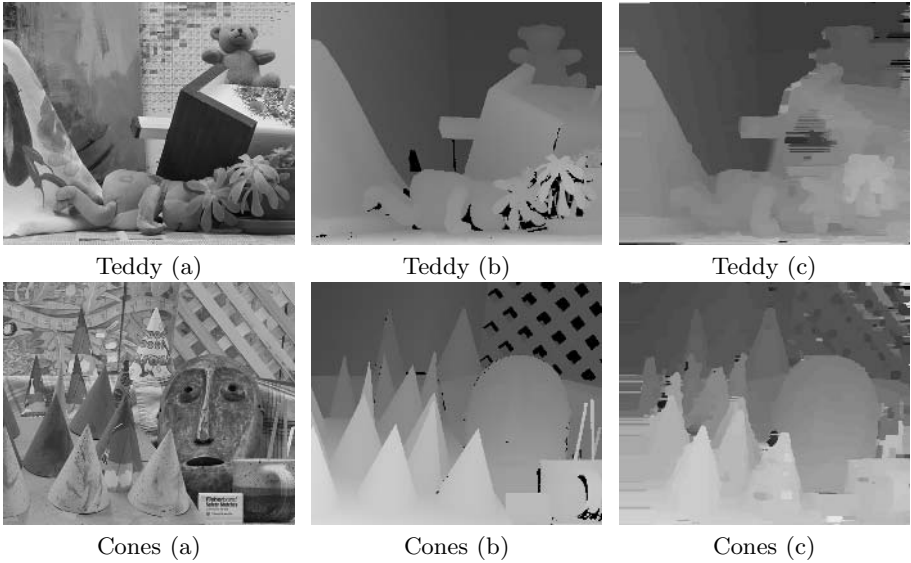


Fig. 4. Disparity maps produced by the proposed method are shown for Teddy and Cones image. (a) Reference image (b) Ground truth (c) Disparity map after energy convergence.

The initial disparity map was obtained by averaging the Birchfield matching cost [2] over 11×3 window. $D_{(x,y)}^{initial}(f)$ denotes the matching cost at point (x, y) with disparity f when obtaining the initial map. w_x and w_y are window size.

$$D_{(x,y)}^{initial}(f) = \frac{1}{w_x \cdot w_y} \sum_{s=x-\frac{w_x}{2}}^{x+\frac{w_x}{2}} \sum_{t=y-\frac{w_y}{2}}^{y+\frac{w_y}{2}} B_{(s,t)}(f), \tag{8}$$

where $B_{(s,t)}(f)$ represents the Birchfield cost at point (x, y) . Above equation was used to find the initial disparity map.

For the energy minimization function, shiftable 3×9 window was used similar to [3].

$$D_{(x,y)}(f) = \min_{k=-\frac{w_y}{2}, 0, \frac{w_y}{2}} \left(\frac{1}{w_y \cdot w_x} \sum_{s=x-\frac{w_x}{2}}^{x+\frac{w_x}{2}} \sum_{t=y+\frac{w_y}{2}}^{y+\frac{w_y}{2}} B_{(s,t+k)}(f) \right). \tag{9}$$

3.2 Evaluation

The computational time is shown in the Table 1. For all calculation Pentium IV 3.4G PC was used. The advantages of proposed method is that the minimization time is independent of the size of disparity range, $O(n)$ iterations. And so,

Table 1. The computational time in seconds. Iterations were performed until the convergence.

	Tsukuba	Sawtooth	venus	map	teddy	cones
Image Size	384x288	434x380	434x383	286x216	450x375	450x375
Max Disparity	15	21	19	28	59	59
time(match cost)	0.40	0.70	0.77	0.41	2.48	2.48
time(DDO)	0.13	0.11	0.18	0.09	0.43	0.40
total time	0.53	0.81	0.95	0.50	2.91	2.88

Table 2. The percentages of error where the disparities are off by more than one on various images; tsukuba, sawtooth, venus, map, teddy, and cones [1]

Techniques	Tsukuba	Sawtooth	venus	map	teddy	cones
Proposed(DDO)	2.26%	1.63%	0.88%	0.35%	14.0%	11.9%
Reliability-DP [8]	1.36%	1.09%	2.35%	0.55%	9.82%	12.9%
Semi Global [9]	3.26%	NA	0.25%	NA	5.14 %	2.77%
Tree DP [14]	1.99%	1.41%	1.41%	1.45%	15.9%	10.0%
4-State DP [1] [5]	4.70%	1.32%	1.53%	0.81%	NA	NA
Sanl. Opt [12]	5.08%	4.06%	9.44%	1.84%	19.9	13.0
DP [12]	4.12%	4.84%	10.10.7%	3.33%	14.0%	10.5%

the majority of time was devoted to aggregating the matching cost. Incremental iteration was performed until the energy function eases to be minimized. Typically, each scanline had less than 100 iterations, and the minimization time itself was semi-real time, but the total computational time is within 1 second for all images, except for Teddy and Cones images. It is worthwhile to note that the program was written with typical C++ without using SIMD (Single Instruction Multiple Data) techniques. We believe that a realtime application of the proposed technique is also possible.

The details of performance is shown on Table 2. The proposed method was compared with DP and the recent variants of DP that emphasized speed of their algorithm. Compared to the scanline DP, disparity discontinuity optimization had clearly better performance, while reliability DP, tree DP, and 4-state DP had a similar performance, although the proposed method ranked highest as in the Table 2 according to [1]. The qualitative performances and ground truth are shown in Fig. 3 and 4.

Unfortunately, in the images with larger maximum disparity, such as Cones and Teddy, the proposed method shows its weakness. The discontinuity optimization rely heavily on the initial conditions. And for a larger disparity range, the good initial disparity map is harder to obtain.

The disparity discontinuity optimization's only fair comparison would DP, since all others are the extension of DP. Reliability DP, tree DP, and 4-state DP can be approached with proposed method. And it is shown that the disparity discontinuity optimization have superior performance than DP, at least for the

test images. Although the speed could be an argumentative case, in all practical programming they are equivalent.

4 Conclusion and Future Work

The initial disparity map and the nature of the proposed algorithm, seek the same goal as the Bobick's GCP; they provide additional constraints during energy minimization. But unlike GCP, the initial disparities can extend, shrink, or be replaced with DDO iteration when a smaller energy function is found, making it more flexible. The proposed algorithm, thus naturally eliminates most of disparity candidates and the unrealistic combinations of disparities. For single iteration, a pixel is assumed to have only two possible disparities; its current value or the one of the disparity across the first discontinuity. The proposed algorithm also prevents the creation of new discontinuities thus allowing fewer combinations. The evaluation from test images shows that the proposed technique is fast with comparable performance with recent variants of DP. For the future study, we believe that the proposed method can be optimized so that it can be done in real or semi-real time. And extending the idea into 2D is also feasible, even though it is indeterminable whether 2D discontinuity model will have same advantages in speed as 1D.

Acknowledgement

This work has been supported in part by the ITRC (Information Technology Research Center) support program of Korean government and IIRC (Image Information Research Center) by Agency of Defense Development, Korea.

References

1. <http://cat.middlebury.edu/stereo/>.
2. S. Birchfield and C. Tomasi. A pixel dissimilarity measure that is insensitive to image sampling. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 20, 1998.
3. A. F. Bobick and S. S. Intille. Large occlusion stereo. *IJCV*, 33, 1999.
4. Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *ICCV*, 1999.
5. A. Criminisi, J. Shotton, and etc. Efficient dense-stereo and novel-view synthesis for gaze manipulation in one-to-one teleconferencing. *Microsoft Technical Report*, 2003.
6. P. F. Felzenszwalb and D. P. Huttenlocher. Efficient belief propagation for early vision. *IEEE Conf. on CVPR*, 2004.
7. D. Geiger, B. Ladendorf, and A. Yuille. Occlusions and binocular stereo. *Proc. European Conf. Computer Vision*, 1992.
8. M. Gong and Y. H. Yang. Near real-time reliable stereo matching using programmable graphics hardware. *CVPR*, 2005.
9. H. Hirschmuller. Accurate and efficient stereo processing by semi-global matching and mutual information. *CVPR*, 2005.

10. H. Ishikawa and D. Geiger. Occlusions, discontinuities, and epipolar lines in stereo. *European Conference on Computer Vision*, 1998.
11. C. Kim, K.M. Lee, B.T. Choi, and S.U. Lee. A dense stereo matching using two-pass dynamic programming with generalized ground control points. *CVPR*, 2005.
12. D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 2002.
13. Jian Sun, Nan-Ning Zheng, and Heung-Yeung Shum. Stereo matching using belief propagation. *IEEE Trans. PAMI*, 25, 2003.
14. O Veksler. Stereo correspondence by dynamic programming on a tree. *CVPR*, 2005.

A New Stereo Matching Model Using Visibility Constraint Based on Disparity Consistency

Ju Yong Chang, Kyoung Mu Lee, and Sang Uk Lee

School of Electrical Eng., ASRI, Seoul National University,
151-600, Seoul, Korea

jangbon@snu.ac.kr, kyoungmu@snu.ac.kr, sanguk@sting.snu.ac.kr

Abstract. There have been many progresses in the stereo matching problem. However, some remaining problems still make stereo matching difficult. Occlusion is one of such problems. In this paper, we propose a new stereo matching model that addresses this problem by using an effective visibility constraint. By considering two images simultaneously, complex geometric configurations regarding the visibility of a pixel becomes simplified, so that the visibility constraint can be modeled as a pairwise MRF. Also since the proposed model enforces the consistency between two disparity maps, the final results become consistent with each other. Belief propagation is employed for the solution of the modeled pairwise MRF. Experimental results on the standard data set demonstrate the effectiveness of our approach.

1 Introduction

Occlusion is one of the major problems in stereo matching, so recently many works have been proposed for handling occlusion [1] [2] [3]. In general, in order to solve the occlusion problem, most researchers tried to impose some constraints to the conventional stereo matching model. And, these constraints can be classified into three classes; ordering constraint, uniqueness constraint, and visibility constraint.

The ordering constraint enforces the order of correspondences to be preserved along the scanlines in both images. Then, a half-occluded region in one scanline can be modeled as pixels that are matched to only one pixel in the other scanline. Conventionally, this ordering constraint has been applied to each scanline independently, and dynamic programming has been used for 1-D optimization problem in each scanline. This inter-scanline independency makes the resultant disparity map have a streaking effect. Recently, Williams et al. [3] imposed the smoothness constraint between scanlines, and applied belief propagation to the resulting 2-D MRF. However, the ordering constraint is not always valid in general setting. Scenes that contain thin foreground objects or narrow holes do not satisfy this property.

The uniqueness constraint [4] enforces a one-to-one correspondence between pixels in two images. Then, a half-occluded region is defined by pixels that do not have corresponding points in the other image. Zitnick and Kanade [5] applied this constraint to an iterative updating scheme for a 3-D match value array in the

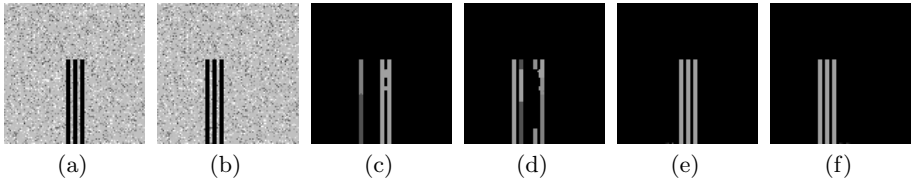


Fig. 1. Synthetic occlusion test example: (a) and (b) are the left and right input images, (c) and (d) are the left and right disparity results produced by symmetric stereo model [1], (e) and (f) are the results by our method

cooperative stereo framework. Kolmogorov and Zabih [6] imposed this constraint on a conventional energy model as a hard constraint of possible assignments and used graph cuts for energy minimization. However, as pointed out in [1] [7], this constraint becomes no longer valid when the scene contains horizontally slanted planes.

The visibility constraint is usually imposed not to enforce the color (intensity) consistency for occluded pixels. However, the occlusion of a pixel depends on many other pixels that can occlude it, so it is not obvious to determine whether a pixel is occluded or not. In the symmetric stereo model, Sun et al. [1] proposed a novel method of inferring the occlusion map in one view by considering the disparity map of the other view. In their work, consistency between the occlusion in one image and the disparity in the other is enforced. By using this constraint, occlusion of scenes that contain thin foreground objects, narrow holes, and horizontally slanted surfaces can be handled appropriately. However, the symmetric stereo model ignores the consistency between two disparity maps. It can be problematic in some scenes as shown in Fig. 1 (a) and (b). This is because the occlusion map in each image is badly inferred by the disparity map of the other image at first iteration. It is obvious that there is lack of consistency between the final two disparity maps obtained by [1] as in Fig. 1 (c) and (d).

In this paper, we propose a new stereo matching model that employs a more effective visibility constraint. If we consider only one image, whether a pixel with a specific disparity is occluded or not depends on many other pixels that can occlude it. However, if both images are considered together, occlusion of a pixel with a specific disparity depends on only one pixel in the other image. So by using the left and right images symmetrically, our new stereo model can be formulated by a pairwise MRF. And the two disparity maps and two occlusion maps are estimated consistently by belief propagation at one step. Fig. 1 (e) and (f) shows the disparity map results by our method for the synthetic test images in Fig.1 (a) and (b).

2 Proposed Stereo Matching Model

Suppose we are given two rectified images. Let L be the set of pixels in the left image, and let R be the set of pixels in the right image. In the binocular stereo,

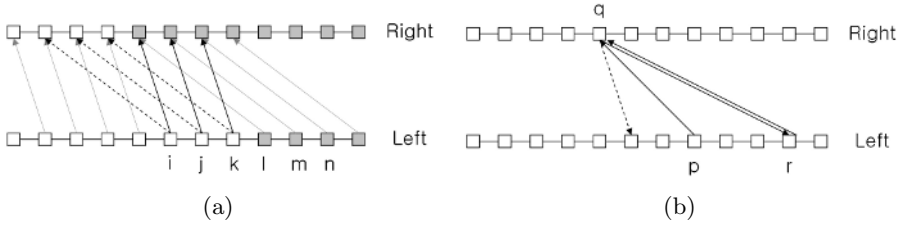


Fig. 2. (a) Solid arrows represent true disparities. But, due to color consistency, pixel i, j, k will have dotted arrow disparities (fattening effect). (b) Suppose that pixel p is matched to pixel q . Then, whether pixel p is occluded or not depends on pixel q . If pixel q is consistently matched to pixel r , pixel p is occluded by pixel r . Pixel q cannot have consistent dotted matching. It is because occluding pixel must have larger disparity than occluded pixel.

our goal is, given stereo image pair I^L and I^R , to compute the disparity map pair f^L and f^R . The disparity map $f^L : I^L \rightarrow F$ is a function that is defined at all pixels of the reference image (in this case, the left image). And, F is a set of discrete disparity values.

As most global stereo algorithms do, in this paper, we adopt the general energy minimization framework where the disparity map is determined by minimizing the following energy function.

$$E(f^L) = \sum_{p \in L} D_p(f_p^L) + \sum_{\{p,q\} \in N} V(f_p^L, f_q^L), \tag{1}$$

where $p \in L$ is a pixel in the left image and $N \subset \{\{p, q\} | p, q \in L \cup R\}$ is the neighborhood system. 4-neighborhood system is used in which pixels $p = (p_x, p_y)$ and $q = (q_x, q_y)$ are neighbors if they are in the same image and $|p_x - q_x| + |p_y - q_y| = 1$. The first term is called as the data term, and measures how well the disparity map fits given images (color-consistency). The second term is the smoothness term which encodes the smoothness constraint for the disparity map.

In this paper, we mean the visibility constraint that the occluded pixels should not be involved in stereo matching by eq. (1). For example, let us consider Fig. 2(a). The solid arrows represent the true disparity values for the left image. We can see that in the right image, points i, j, k of the white object are occluded by points l, m, n of the grey object. Unless occluded pixels i, j, k are treated separately, the data terms of those pixels may bring false matches. For the case of Fig. 2(a), because of the color (intensity) consistency and the smoothness constraint, pixels i, j, k have disparities corresponding to dotted arrows, causing the fattening effect.

Therefore, modification of the data term is necessary for handling the occluded pixels properly. Imposing zero or some small constant to the data terms corresponding to the occluded pixels is a common method to exclude the effect of occluded pixels in matching. Detecting whether a pixel is occluded or not is

another important problem, and generally it depends on the disparity of that pixel as well as the disparities of all pixels that may occlude that pixel. Without loss of generality, let us assume $F = \{0, 1, \dots, n - 1\}$. Then, pixel $p = (p_x, p_y)$ in the left image can be occluded by $(n - 2)$ pixels, that is, $p + 1 = (p_x + 1, p_y)$, $p + 2 = (p_x + 2, p_y), \dots, p + n - 2 = (p_x + n - 2, p_y)$. Thus, a function indicating whether a pixel p is occluded or not can be represented by the visibility function, $Vis_p(f_p, f_{p+1}, f_{p+2}, \dots, f_{p+n-2})$. Suppose that Vis_p is defined by 0 if pixel p is occluded, and 1 otherwise. Then, the modified data term that can eliminate the influence of the occluded pixels can be defined by

$$\sum_p Vis_p(f_p, f_{p+1}, f_{p+2}, \dots, f_{p+n-2}) D_p(f_p). \quad (2)$$

Recently, most successful global stereo algorithms use the graph cuts [6] [8] [9] [10] or belief propagation [1] [11] for minimizing energy. However, applying such algorithms directly to the minimization of our modified energy function is not easy. This is because the modified energy does not satisfy the regularity condition that is the necessary and sufficient condition for using the graph cuts [12]. And also the modified energy function is no longer in pairwise form, so that it cannot be modeled by pairwise MRF. Actually, the data term in eq. (2) corresponds to the Gibbs energy of $(n - 1)$ -wise MRF. Hence, belief propagation is not directly applicable to the minimization of our modified energy function.

Note that according to the result in [13], any MRF with higher order cliques can, in principle, be converted to a pairwise MRF defined on an augmented graph. And the augmented graph can be obtained by suitable clustering of nodes into large nodes. However, in the case of our modified energy, since the number of possible states that the clustered nodes can have is too large, applying this method is not practical too.

Therefore we devise a new stereo energy model that is practically tractable, and it is inspired by the following observation. Let us consider the situation in Fig. 2(b). Assume that pixel p in the left image is matched to pixel q in the right image. Then, how can we know whether pixel p is occluded or not? Let us introduce a new auxiliary variable $c^L : I^L \rightarrow \{true, false\}$ defined at all pixels in the left image. This new boolean variable represents whether the disparity of a pixel is consistent or not. That is, in Fig. 2(b), if pixel q has the solid arrow disparity and its boolean variable is *true*, pixel r also must have the solid arrow disparity (consistency). However, if the boolean variable of pixel q is *false*, the disparity of pixel r cannot be the solid arrow (inconsistency). Now whether pixel p is occluded or not is determined by variables of pixel q , that is, its disparity and boolean variable. Suppose that pixel q has the solid arrow disparity and *true*. Then, because not only pixel p but also pixel r are matched to pixel q , we can think that pixel p is occluded by pixel r . By using this idea, we can impose the visibility constraint on the existing energy function by pairwise form. Detailed equations will be presented in Section 2.1.

Note that the disparity maps and new boolean variable maps of both images must satisfy several restricted configurations. For example, in Fig. 2(b), let us assume that pixel p has solid arrow disparity and *false*. Then, the boolean variable of pixel q have to be *true*. It is because for pixel p to be occluded, one non-occluded pixel in the left image must be consistently matched to pixel q in the right image. And the disparity of pixel q have to be larger than that of pixel p . It is because the occluding object must be more closely than the occluded object, and the depth is inversely proportional to the disparity. These restrictions of possible configurations for disparity maps and boolean variable maps in two images enforce consistency between two disparity maps. This is a new feature that is distinct from the symmetric stereo matching model in [1]. In order to constrain possible configurations, the new model should includes an additional consistency energy term, and more complete description about this will be given in Section 2.1.

2.1 A New Energy Function

Now our new energy function can be formulated by

$$\begin{aligned}
 E(c^L, f^L, c^R, f^R) = & Data_L(c^L, f^L) + Data_R(c^R, f^R) \\
 & + Smooth_L(c^L, f^L) + Smooth_R(c^R, f^R) \\
 & + Vis(c^L, f^L, c^R, f^R) + Con(c^L, f^L, c^R, f^R).
 \end{aligned}
 \tag{3}$$

The first and second terms are the data terms, and the third and fourth terms are the smoothness terms. The fifth term is the visibility term where the visibility constraint is encoded, and the last term is the consistency term that enforces the consistency between the left and right disparity maps.

The data term imposes color (intensity)-consistency. For the left image,

$$Data_L(c^L, f^L) = \sum_{p \in L} D_p(f_p),
 \tag{4}$$

where $D_p(f_p)$ measures the degree of color (intensity)-consistency between pixel p and its corresponding pixel when the disparity of pixel p is f_p . For $D_p(f_p)$, we can use various matching functions such as SD (Squared Difference), AD (Absolute Difference), and the Birchfield measure [14]. In this paper, according to [1], the following truncated L_1 norm function that is robust to noise and outliers is used:

$$D_p(f_p) = -\ln((1 - e_d)\exp(-\|I_p^L - I_q^R\|/\sigma_d) + e_d),
 \tag{5}$$

where pixel $q \in R$ is the matched point of pixel p whose disparity is f_p . The data term for the right image, $Data_R(c^R, f^R)$ can be defined symmetrically.

The smoothness term encodes the smoothness constraint. For the left image, we define the smoothness term as follows:

$$Smooth_L(c^L, f^L) = \sum_{\{p,q\} \in N} V_{\{p,q\}}(f_p^L, f_q^L).
 \tag{6}$$

For $V_{\{p,q\}}(f_p, f_q)$, we use following robust L_1 distance:

$$V_{\{p,q\}}(f_p, f_q) = \min(\lambda|f_p - f_q|, \lambda \cdot \mu), \tag{7}$$

where λ is the rate of increase in the cost, and μ controls the limit of the cost. Employing this robust function has several advantages. First, as a discontinuity preserving constraint, this function can recover discontinuous features of true disparity map very well. Furthermore, according to [15], the implementation of belief propagation for minimizing the energy function containing this smoothness term can be done efficiently by using distance transform. The smoothness term for the right image, $Smooth_R(c^R, f^R)$ can be defined symmetrically. And, we can observe that for the data terms and smoothness terms, the boolean variables c^L , c^R have no influence on the energy.

The visibility term where the visibility constraint is encoded can be represented by sum of two terms for occlusion in the left and right images as follows:

$$Vis(c^L, f^L, c^R, f^R) = \sum_{p \in L, \{p,q\} \in \overline{N}} \overline{V}_{\{p,q\}}^L(c_p^L, f_p^L, c_q^R, f_q^R) + \sum_{p \in R, \{p,q\} \in \overline{N}} \overline{V}_{\{p,q\}}^R(c_p^R, f_p^R, c_q^L, f_q^L). \tag{8}$$

$\overline{N} \subset \{\{p, q\} | p, q \in L \cup R\}$ is a new neighborhood system to describe the interactions between two images. In this new neighborhood system, a neighborhood of a pixel p is defined by all pixels in the other image that pixel p can correspond to. For example, if we suppose the possible disparity range of the left image to be $\{0, 1, \dots, n-1\}$, then the neighbors of pixel (p_x, p_y) in the left image become pixels $(p_x, p_y), (p_x - 1, p_y), \dots, (p_x - (n - 1), p_y)$ in the right image. $\overline{V}_{\{p,q\}}^L(c_p^L, f_p^L, c_q^R, f_q^R)$ is an occlusion function of pixel p in the left image by pixel q in the right image, defined by

$$\overline{V}_{\{p,q\}}^L(c_p^L, f_p^L, c_q^R, f_q^R) = \begin{cases} -D_p(f_p^L) + k, & \text{if } p_x + f_p^L = q_x \wedge c_q^R = true \wedge f_p^L < f_q^R; \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

As shown in Fig. 2(b), when pixel p in the left image corresponds to pixel q in the right image ($p_x + f_p^L = q_x$), if pixel q is consistent ($c_q^R = true$) and disparity of pixel q is larger than that of pixel p ($f_p^L < f_q^R$), then pixel p is occluded by pixel q . Thus, we subtract the data term of pixel p for its disparity f_p^L from the energy function, and instead add a small constant k to it. Otherwise, the visibility term has no effect on the energy function. $\overline{V}_{\{p,q\}}^R(c_p^R, f_p^R, c_q^L, f_q^L)$ can be defined analogously.

The last term, that is, the consistency term enforces the consistency between the left and right disparity maps. This term can be written by

$$Con(c^L, f^L, c^R, f^R) = \sum_{\{p,q\} \in \overline{N}} C_{\{p,q\}}(c_p^L, f_p^L, c_q^R, f_q^R), \tag{10}$$

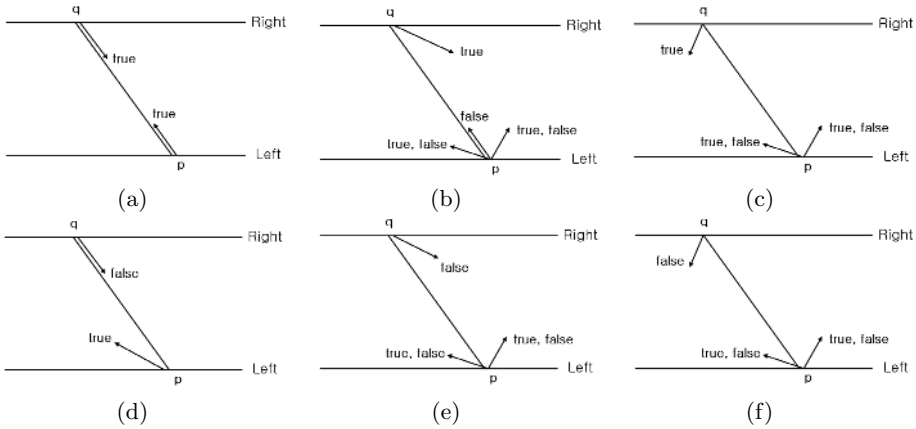


Fig. 3. Assume that pixel p is in the left image, pixel q is in the right image, and $p_x + d = q_x$. Then, valid configurations of two pixels classified by variable of right image pixel q : (a) $c_q^R = true, f_q^R = d$; (b) $c_q^R = true, f_q^R > d$; (c) $c_q^R = true, f_q^R < d$; (d) $c_q^R = false, f_q^R = d$; (e) $c_q^R = false, f_q^R > d$; (f) $c_q^R = false, f_q^R < d$.

where $C_{\{p,q\}}(c_p^L, f_p^L, c_q^R, f_q^R)$ constrains possible configurations of two pixels p and q that belong to the new neighborhood system \overline{N} , and has the following form:

$$C_{\{p,q\}}(c_p^L, f_p^L, c_q^R, f_q^R) = \begin{cases} 0, & \text{if } \{c_p^L, f_p^L, c_q^R, f_q^R\} \text{ is valid configuration;} \\ \infty, & \text{otherwise.} \end{cases} \quad (11)$$

This function inhibits invalid configurations of two pixels p and q by imposing infinity on the energy function. Since the complete description of all the valid configurations is too complicated, instead we present a pictorial description as in Fig. 3. In Fig. 3, we assume that pixel p is in the left image, pixel q is in the right image, and $p_x + d = q_x$. Then, valid configurations of two pixels can be classified by variables of pixel q in the right image, that is, c_q^R and f_q^R . Firstly, if $c_q^R = true$ and $f_q^R = d$, it means that pixel q is consistently matched to pixel p , so pixel p also have to be consistently matched to pixel q ($c_p^L = true, f_p^L = d$). Other values of c_p^L and f_p^L are invalid. Secondly, if $c_q^R = true$ and $f_q^R > d$, then, only $c_p^L = true$ and $f_p^L = d$ is forbidden. But, when $c_q^R = true$ and $f_q^R < d$, then $c_p^L = false$ and $f_p^L = d$ is additionally forbidden. Otherwise, farther pixel q occludes closer pixel p . And, if $c_q^R = false$ and $f_q^R = d$, it means that pixel q is occluded by pixel p , so pixel p must be consistently matched and have larger disparity than pixel q ($c_p^L = true, f_p^L > d$). Finally, if $c_q^R = false$ and $f_q^R \neq d$, then pixel p cannot have $c_p^L = true, f_p^L = d$. Moreover pixel p cannot have $c_p^L = false, f_p^L = d$, too. It is because an occluded pixel ($c_p^L = false$) cannot correspond to another occluded pixel ($c_q^R = false$).

3 Optimization Using Belief Propagation

The new energy model in eq. (3) can be written in the following pairwise form:

$$E(l) = \sum_{p \in P} \theta_p(l_p) + \sum_{\{p,q\} \in \tilde{N}} \theta_{\{p,q\}}(l_p, l_q), \quad (12)$$

where l_p is $\{c_p, f_p\}$, P is $L \cup R$, \tilde{N} is $N \cup \bar{N}$, $\theta_p(l_p)$ is a unary data penalty function, and $\theta_{\{p,q\}}(l_p, l_q)$ is a pairwise interaction potential. we note that similar form of energy function has been derived in the context of MRF [16], and applied to many early vision problems. And a minimum of this energy corresponds to a maximum a-posteriori (MAP) labeling.

In general, minimizing a pairwise energy function is an NP-hard problem, so researchers have focused on approximate minimization algorithms. Two successful algorithms are graph cuts [8] [12] [17] [18] and belief propagation [11] [15] [19] [20]. To our knowledge, graph cuts are known to be able to reach lower energy than belief propagation [21]. But since our energy model does not satisfy the regularity condition, we cannot apply graph cuts to it. Therefore, in this paper, we use belief propagation to minimize the proposed energy.

4 Experimental Results

In our algorithm, the parameter λ is automatically determined by using a similarity based technique as in [1], in which the similarity between pixels is computed via the Kullback-Leiber (KL) divergence, and λ of an image is set to be proportional to the average similarity in the image. We used 0.25 as the proportional constant. Analogously, the parameter μ inside an image is determined differently according to the color (intensity) similarity between pixels as follows:

$$\mu_{\{p,q\}} = \begin{cases} 2, & \text{if } p \text{ and } q \text{ belong to different segments;} \\ 3.8, & \text{otherwise.} \end{cases} \quad (13)$$

The mean-shift algorithm was used for classifying pixels into segments [22]. Other parameters are fixed as $\sigma_d = 4.0$, $e_d = 0.01$, and $k = 3.0$.

We evaluate the proposed algorithm using four standard datasets, Tsukuba, Sawtooth, Venus, and Map in [23]. In [23], a pixel is considered erroneous if its absolute disparity error is greater than one. And the percentages of bad pixels are computed in all region (all), textureless region (untex.), and discontinuity region (disc.). Notice that only nonoccluded pixels are considered in these all computations.

Fig. 4 shows the disparity maps and bad pixel results by our method. We can observe that very good performances have been achieved in the occluded region as well as other areas. Table 1 presents the overall performance of our algorithm and quantitative comparison with other algorithms. Our algorithm ranked the third out of 36 algorithms, and has little difference from the top-ranked algorithm. Note that our method produced the best performances (in 'all'

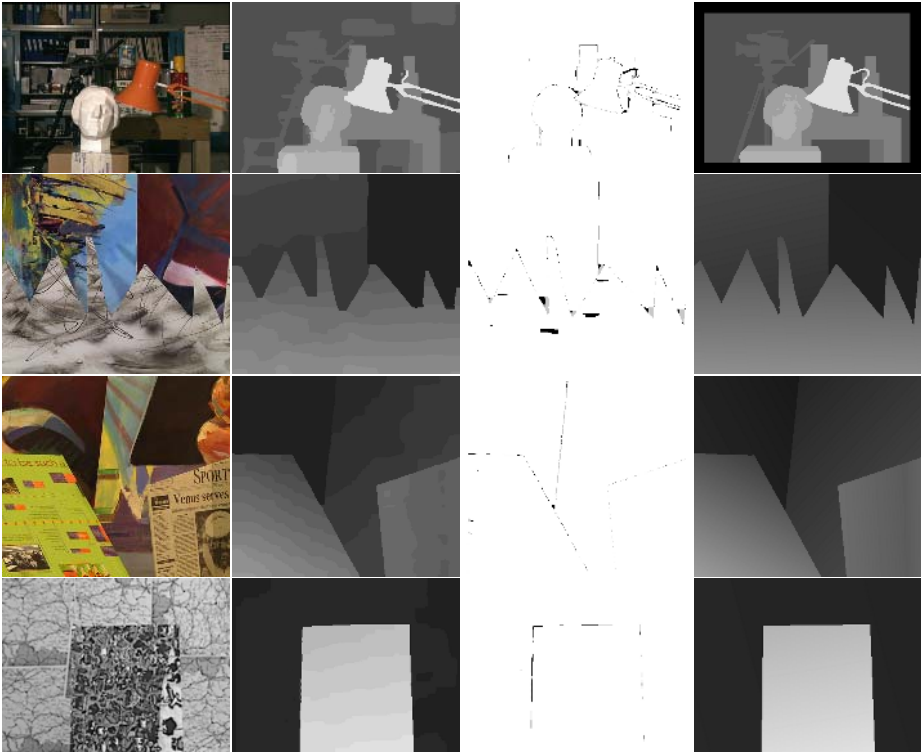


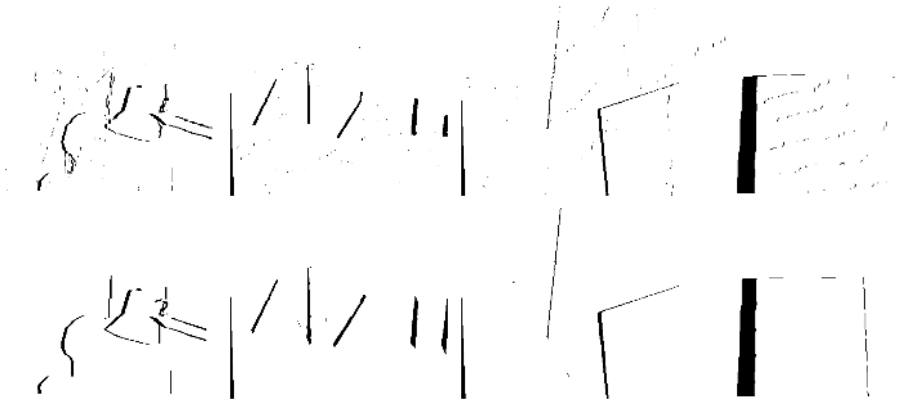
Fig. 4. Results of the proposed algorithm on Middlebury datasets. From top to down order: Tsukuba, Sawtooth, Venus, Map. From left to right order: reference images, extracted disparity maps, bad pixel images, and the ground truth disparity maps.

and 'untext.' regions) for the Tsukuba images. However, for other test images, especially for the Sawtooth images, there is still some rooms for improvement. As can be seen from the bad pixels result of the Sawtooth images in Fig. 4, errors are occurred mainly in the inner region of the slanted plane. So, we expect that robust plane fitting and subsequently using the fitting result as soft constraint in [1] can reduce those errors significantly.

To demonstrate the performance of the proposed stereo model on occlusion, we presents our occlusion results in Fig. 5 and compare them quantitatively with those of the other state of the art algorithms in Table 2. The percentages of false positives, false negatives, and bad pixels in the occluded region are listed for the Tsukuba images. For the false positive and false negative results of other algorithms, we referred to [1] [24]. And we downloaded the disparity results of other algorithms from the Middlebury stereo vision page (<http://www.middlebury.edu/stereo>), and computed the percentages of bad pixels over the occluded regions. We can see that our algorithm ranked 2nd in all categories, and produced excellent results.

Table 1. Evaluation table of different stereo algorithms

Algorithms	Tsukuba			Sawtooth			Venus			Map	
	all	untex.	disc.	all	untex.	disc.	all	untex.	disc.	all	disc.
Sym.BP [1]	0.97	0.28	5.45	0.19	0.00	2.09	0.16	0.02	2.77	0.16	2.20
Patch-based [24]	0.88	0.19	4.95	0.29	0.00	3.23	0.09	0.02	1.50	0.30	4.08
Our method	0.87	0.14	5.07	0.80	0.04	5.11	0.15	0.02	2.35	0.29	4.05
Seg.GC [10]	1.23	0.29	6.94	0.30	0.00	3.24	0.08	0.01	1.39	1.49	15.46
Graph+segm.	1.39	0.28	7.17	0.25	0.00	2.56	0.11	0.02	2.04	2.35	20.87
Segm.+glob.vis.	1.30	0.48	7.50	0.20	0.00	2.30	0.79	0.81	6.37	1.63	16.07
Layered [25]	1.58	1.06	8.82	0.34	0.00	3.35	1.52	2.96	2.62	0.37	5.24
Belief prop. [11]	1.15	0.42	6.31	0.98	0.30	4.83	1.00	0.76	9.13	0.84	5.27
Region-Progress.	1.44	0.55	8.18	0.24	0.00	2.64	0.99	1.37	6.40	1.49	17.11
2-pass DP	1.53	0.66	8.25	0.61	0.02	5.25	0.94	0.95	5.72	0.70	9.32

**Fig. 5.** Occlusion results on Middlebury datasets. From top to down order: extracted occlusion maps, and the ground truth occlusion maps. From left to right order: Tsukuba, Sawtooth, Venus, Map.**Table 2.** Occlusion evaluation for Tsukuba dataset

Algorithms	False positives		False negatives		Errors in occl.	
	rate	rank	rate	rank	rate	rank
Our method	0.87	2	29.58	2	22.10	2
Sym.BP [1]	0.7	1	29.9	3	31.62	4
Patch-based [24]	1.05	3	30.16	4	39.41	5
Seg.GC [10]	1.19	4	32.51	5	18.95	1
Layered [25]	2.28	5	25.42	1	29.36	3

5 Conclusions

In this paper, we presented a new stereo matching model using an effective visibility constraint. The main contributions of this paper are as follows. By

using two input images simultaneously, the visibility constraint is modelled by a pairwise MRF. And by enforcing the consistency between two disparity maps, consistent and more accurate results could be obtained. Finally, our results on real data demonstrated the effectiveness of proposed method on occluded areas.

Acknowledgement

This work has been supported in part by the ITRC (Information Technology Research Center) support program of Korean government and IIRC (Image Information Research Center) by Agency of Defense Development, Korea.

References

1. Sun, J., Li, Y., Kang, S.B., Shum, H.Y.: Symmetric stereo matching for occlusion handling. In: CVPR05. (2005) II: 399–406
2. Wei, Y., Quan, L.: Asymmetrical occlusion handling using graph cut for multi-view stereo. In: CVPR05. (2005) II: 902–909
3. Williams, O., Isard, M., MacCormick, J.: Estimating disparity and occlusions in stereo video sequences. In: CVPR05. (2005) II: 250–257
4. Marr, D., Poggio, T.A.: Cooperative computation of stereo disparity. *Science* **194** (1976) 283–287
5. Zitnick, C.L., Kanade, T.: A cooperative algorithm for stereo matching and occlusion detection. *PAMI* **22** (2000) 675–684
6. Kolmogorov, V., Zabih, R.: Computing visual correspondence with occlusions via graph cuts. In: CVPR04. (2004) I: 261–268
7. Ogale, A.S., Aloimonos, Y.: Stereo correspondence with slanted surface: critical implication of horizontal slant. In: CVPR04. (2004) I: 568–573
8. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. *PAMI* **23** (2001) 1222–1239
9. Kolmogorov, V., Zabih, R.: Multi-camera scene reconstruction via graph cuts. In: ECCV02. (2002) III: 82–96
10. Hong, L., Chen, G.: Segment-based stereo matching using graph cuts. In: CVPR04. (2004) I: 74–81
11. Sun, J., Zheng, N., Shum, H.: Stereo matching using belief propagation. *PAMI* **25** (2003) 787–800
12. Kolmogorov, V., Zabih, R.: What energy functions can be minimized via graph cuts. *PAMI* **26** (2004) 147–159
13. Weiss, Y., Freeman, W.T.: On the optimality of solutions of the max-product belief propagation algorithm in arbitrary graphs. *IEEE Transactions on Information Theory* **47** (2001) 723–735
14. Birchfield, S., Tomasi, C.: A pixel dissimilarity measure that is insensitive to image sampling. *PAMI* **20** (1998) 401–406
15. Felzenszwalb, P.R., Huttenlocher, D.P.: Efficient belief propagation for early vision. In: CVPR04. (2004) I: 261–268
16. Geman, S., Geman, D.: Stochastic relaxation, gibbs distributions, and the bayesian restoration of images. *PAMI* **6** (1984) 721–741
17. Greig, D., Porteous, B., Seheult, A.: Exact maximum a posteriori estimation for binary images. *Journal of the Royal Statistical Society* **51** (1989) 271–279

18. Ishikawa, H.: Exact optimization for markov random fields with convex priors. *PAMI* **25** (2003) 1333–1336
19. Pearl, J., ed.: Probabilistic reasoning in intelligent systems: networks of plausible inference. Morgan Kaufmann Publishers Inc. (1988)
20. Yedidia, J.S., Freeman, W.T., Weiss, Y.: Generalized belief propagation. In: *Advances in Neural Information Processing Systems*. (2000) 689–695
21. Tappen, F., Freeman, W.T.: Comparison of graph cuts with belief propagation for stereo, using identical mrf parameters. In: *ICCV03*. (2003) II: 900–906
22. Comaniciu, D., Meer, P.: Mean shift: A robust approach toward feature space analysis. *PAMI* **23** (2001) 603–619
23. Scharstein, D., Szeliski, R.: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV* **47** (2002) 7–42
24. Deng, Y., Yang, Q., Lin, X., Tang, X.: A symmetric patch-based correspondence model for occlusion handling. In: *ICCV05*. (2005)
25. Lin, M., Tomasi, C.: Surfaces with oclusions from layered stereo. *PAMI* **26** (2004) 1073–1078

Refine Stereo Correspondence Using Bayesian Network and Dynamic Programming on a Color Based Minimal Span Tree

Naveed I Rao¹, Huijun Di¹, and GuangYou Xu²

¹ Pervasive Computing Lab, Institute of Human Computer Interaction Department of Computer Engineering, Tsinghua University, Beijing, China

{naveed03, dhj98}@mails.tsinghua.edu.cn

² {xgy-dcs}@mail.tsinghua.edu.cn

Abstract. Stereo correspondence is one of the basic and most important problems in computer vision. For better correspondence, we need to determine the occlusion. Recently dynamic programming on a minimal span tree (mst) structure is used to search for correspondence. We have extended this idea. First, mst is generated directly based on the color information in the image instead of converting the color image into a gray scale. Second, have treated this mst as a Bayesian Network. Novelty is attained by considering local variances of the disparity and intensity differences in the conditional Gaussians as unobserved random parameters. These parameters are iteratively inferred by alternate estimation along the tree given a current disparity map. It is followed by dynamic programming estimation of the map given the current variance estimates thus reducing the overall occlusion. We evaluate our algorithm on the benchmark Middlebury database. The results are promising for modeling occlusion in early vision problems.

1 Introduction

Occlusion is one of the major challenges in stereo vision. In stereo, occlusion corresponds to a specific situation, that some points in the scene are visible to one camera but not the other due to the scene and camera geometries [12]. But in this work, we refer to the occlusion as if a point in left image (L) could not find its correspondence in the right image (R). Detection of these occluded pixel is ambiguous, so prior constraints need to be imposed e.g. ordering constraint [4] is exploited in dynamic programming framework [11] as it reduces the search space. These occluded pixels are excluded based on the threshold [14],[3] and the scene is assumed as free of occlusion. This naturally entails the piecewise smoothness of the recovered stereo correspondence map or disparity map [4].

Motivation for this work is to develop a stereo corresponding algorithm, which can model occlusion and later can reduce it with possible low computational cost. These goals are achieved by modeling occlusion using bayesian network and achieved low computation cost by utilizing dynamic programming (DP) on

tree network. The bayesian methods (e.g.,[8],[9],[6], [1],[2]) globally model discontinuities and occlusion. These methods can be classified into two categories [13] based on their computational model i.e. dynamic programming-based and MRFs-based. Keeping in view the scope of this work, we will cover the former category in detail.

Geiger et al. [8] and Ishikawa, Geiger [9] derived an occlusion process and a disparity field from a matching process. The matching process is transformed to a path-finding problem by assuming order constraint and uniqueness constraint, where the global optimum is obtained by dynamic programming. Belhumeur [1] used a simplified relationship between disparity and occlusion to solve scan line matching by dynamic programming and by defining a set of priors from a simple scene to a complex scene. Contrary to above where a piecewise-smooth constraint is imposed, Cox et al.[6] and Bobick and Intille [2] did not require the smoothing prior. They assumed a normal distribution of corresponding features and a fixed cost for occlusion, and using only the occlusion constraint and ordering constraints, they proposed a dynamic programming solution. The work of Bobick and Intille focused on reducing the sensitivity to occlusion cost and the computation complexity of Cox's method, by incorporating the ground control points constraint. These dynamic programming methods are employed with the assumption of same occlusion cost in each scan line. Ignoring the dependence between scan lines, results in the characteristic streaking in the disparity maps. State of art results can be achieved by 2D, but at the cost of time.

In comparison, our approach is much simpler, but also much more efficient. To seek global optimum with linear degree of search, we have taken advantage of mini-mal span tree (mst) network. The contribution by this paper is to treat, first time in known literature, mst as a Bayesian Network. To deal with occlusion and for controlling smoothness in disparity space, two random parameters are introduced in the network. Instead of exact inference, posterior distribution is approximated by computing Helmholtz free energy using EM on each node of tree. These energies are minimized by using dynamic programming. In this way we have fused mst, bayesian network and dynamic programming into one method to find the optimal disparity in the image. The algorithm is tested on Middlebury stereo database and the results in comparison to state of art algorithms are promising.

The paper is organized as follows: Section II explains details of general framework of the problem along extraction of mst based on the color information instead of gray scale values as been done by recent works [14],occlusion modeling, computation of free energies and minimization of energies using DP. Comparison with other algorithms and experimental results are shown in Section III. Conclusion forms the last section.

2 Occlusion Modeling

In this section, we will first introduce general framework of problem, later the concept of treating mst as a bayesian network is explained. Local variances of the

disparity and intensity differences in the conditional gaussians are introduced as unobserved random parameters. In order to approximate the posterior probability Helmholtz Free Energy is calculated at every node and is explained in next subsection. These parameters are iteratively inferred by applying expectation maximization while the current disparity map is given. A new optimal disparity map is attained by minimizing energy through dynamic programming on tree and forms the last subsection.

2.1 General Framework

In this subsection, we have explained the general framework of our problem. The notations used are borrowed from [14]. Let $G(V, E)$ be a grid connected graph with vertices V and edges E . All pixels of the left image form the vertices in V . For the edges E , every pixel p is connected with his 4-connected neighbors. We want to convert this into a tree graph $G'(V, E')$ by choosing the most valueable edge E' of each pixel. The definition of most valueable edges out of 4-connected edges, remained under investigation. We exploited the fact, that disparity discontinuities align with intensity discontinuities. It means that if the neighboring pixels i and j have similar intensity values i.e. $Z(i)$ and $Z(j)$, then they are more likely to have same disparity a priori. The gray intensity information provided by the left image is used to assign different weights to edges in $G(V, E)$ [14]. The intensity information provided in shape of gray level is not sufficient to decide the strength of the edges in between pixels. The conversion of color pixel into gray scale is as mapping from many to one problem. In this way, the distance achieved between pixels is not a true distance. Instead in our case we have we have used the distance among the pixels based on the definition of color components.

$$d_2(v_1, v_2) > 0 \Leftrightarrow \begin{bmatrix} r_1 \\ g_1 \\ b_1 \end{bmatrix} - \begin{bmatrix} r_2 \\ g_2 \\ b_2 \end{bmatrix} \quad (1)$$

$G(V, E)$ is converted into tree graph i.e. mst by using standard Kruskal's Algorithm[5]. Since edge weights are integers in small range, the edge sorting can be performed in linear time, therefore the mst can be computed in basically linear time. Later, experimental results have proved that mst extracted using the color information, provides much better results. These results does not depend upon the specific data but represents that color images provide better smoothness as compared to the gray scale images.

2.2 Formulation of Problem in Bayesian Approach

Bayesian networks are used for modeling uncertainties in the parametric form. Fig.1 shows the Bayesian network (i.e. based on the mst network) used for the current problem. Node d_i is an optimal displacement for pixel i in L Image. $\tau_{i,j}$ is variance in disparity between child i and parent j nodes and controls the smoothness in disparity space. Z_i represents the intensity of each pixel and σ_i is

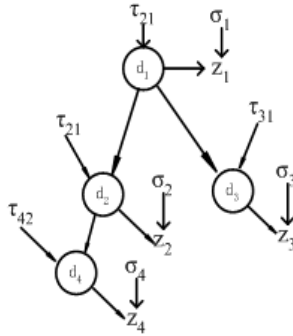


Fig. 1. Occlusion modeling using Bayesian network: τ is local variance of the disparity and σ is variance of the intensity difference

variance between $Z_L(i)$ and $Z_R(i + d)$ respectively. In the network $d_i, j, \tau_{i,j}, \sigma_i$ are hidden and Z_i is visible node. Based on the model, joint distribution is given by the following product of distributions model,

$$\begin{aligned}
 p(\sigma_i, \tau_{i,j}, d_j, Z_i) = & \left\{ \left(\prod_{i=1}^n p(\sigma_i) \right) \cdot \left(\prod_{i=1}^n p(\tau_{i,j}) p(d_{root}) \right) \right. \\
 & \left. \left(\prod_{(i,j) \in E}^{n-1} p(d_i | d_i, \tau_{i,j}) \right) \right\} \cdot \left\{ \prod_{i=1}^n p(z_i | d_i, \sigma_i) \right\}
 \end{aligned} \tag{2}$$

Where n are total numbers of nodes in the tree and d_{root} represents the root node (since root does not have parent so for ease of understanding it is expressed as a separate term). Exact inference requires the computation of posterior distribution, over all hidden variables given the visible, which is often intractable. So we turn to approximation methods which estimate a simple distribution that is close to the correct posterior distribution.

2.3 Computation of Helmholtz Free Energies

The idea is to approximate the true posterior distribution $p(h|v)$ by a simpler distribution $Q(h)$ while h, v represents hidden and visible nodes respectively. A natural choice for a measure of similarity between the two distributions is the relative entropy (a.k.a. Kullback-Leibler divergence), which can be formulated as [7]

$$F(Q, P) = \int_h Q(h) \log Q(h) - \int_h Q(h) \log p(h, v) \tag{3}$$

Where $F(Q, P)$ is Helmholtz free energy or the Gibbs free energy, or just the free energy between Q and P . Intuitively, minimum of the energy can be achieved only once $p(h, v)$ will have same value as $Q(h)$. For ease in display and

understanding above expression is replicated in two parts. $F(Q, P) = T_1 - T_2$. Point inference searching techniques looks for a single configuration h_{approx} of the hidden variable. By befitting a problem as mentioned above, transforms it to minimizing cost problem. We have also adopted the same way. σ and τ are estimated by point estimation and the Q -Distribution for the entire model is:

$$Q(h) = \left(\prod_{i=1}^n \delta(\sigma_i - \hat{\sigma}_i) \right) \cdot \left(\prod_{i=1}^n \delta(\tau_{i,j} - \hat{\tau}_{i,j}) \right) \tag{4}$$

δ is a dirac delta function and for distribution $Q(h)$ is an infinite spike of the density at \hat{h} . For detail properties of δ function please see [7]. By replacing Eq.4 in $T1$ of Eq.3 and after rearranging

$$\begin{aligned} T_1 = & \int_{\tau} Q(\hat{\tau}) \cdot \left[\int_{d_{root}} Q(d_{root}|\hat{\tau}) \cdot \log Q(d_{root}|\hat{\tau}) + \left\{ \sum_{i_h \in C d_{i_{n-1}}} \int_{d_{i1}} Q(d_{i1}|d_{root}, \hat{\tau}) \cdot \right. \right. \\ & \log Q(d_{i1}|d_{root}, \hat{\tau}) + [\dots + \sum_{i_h \in C d_{i_{n-1}}} \int_{d_{in}} Q(d_{in}|d_{n-1}, \hat{\tau}) \cdot \\ & \left. \left. \log Q(d_{in}|d_{n-1}, \hat{\tau})] \dots \right\} + \int_{\sigma_i} Q(\hat{\sigma}_i) \log Q(\hat{\sigma}_i) + \int_{\tau} Q(\hat{\tau}_{i,j}) \log Q(\hat{\tau}_{i,j}) \right] \end{aligned} \tag{5}$$

The last two terms are the entropy of the delta-functions i.e. H_{δ} , and is constant w.r.t. the optimization. Since we intend to use tree structure, and want to apply recursive programming so by rearranging $T1$ in recursive fashion.

$$F_{Q,Q}(d_i, d_j) = \int_{d_i} \left\{ Q(d_i|d_j, \hat{\tau}_{i,j}) \cdot \log Q(d_i|d_j, \hat{\tau}_{i,j}) + \sum_{k \in C} F_{QQ}(d_k, d_i) \right\} \tag{6}$$

Where, C is set of child nodes. For every node i , Q is a matrix of order $l \times l$, for all possible values of d_i and d_j where l is total search area for every pixel. For ease of understanding, Eq.2 is mainly divided into two parenthesis i.e. $p1$ and $p2$. By going through the same procedure as above, for $T2$ in Eq. 3, below are two outcomes of equations $p1$ and $p2$ from Eq.2 respectively.

$$\begin{aligned} F_{QPQ}(d_i, d_j) = & \int_{d_i} \left\{ Q(d_i|d_j, \hat{\tau}_{i,j}) + \sum_{k \in C} F_{QPQ}(d_k, d_i) \right\} + \\ & \int_{\sigma_i} Q(\hat{\sigma}_i) \log Q(\hat{\sigma}_i) + \int_{\tau} Q(\hat{\tau}_{i,j}) \log Q(\hat{\tau}_{i,j}) \end{aligned} \tag{7}$$

$$F_{QPZ}(d_i, d_j) = \int_{d_i} \left\{ Q(d_i|d_j, \hat{\tau}_{i,j}) \cdot \log p(Z_i|d_j, \hat{\tau}_{i,j}) + \sum_{k \in C} F_{QPZ}(d_k, d_i) \right\} \tag{8}$$

The generalized recursive expression for the total free energy can be written as

$$F(Q, P) = \left\{ F_{QQ}(d_{root}, d_{-1}) - F_{QPQ}(d_{root}, d_{-1}) - F_{QPZ}(d_{root}, d_{-1}) + \mathbf{H}_{\delta} \right\} \tag{9}$$

d_{-1} is a dummy node i.e. root node of the root and is shown for simplicity in expression. All terms related with the entropy of the delta function are ignored as they do not take part in optimization. The total free energy can be viewed as the summation of energies at all nodes over the tree.

2.4 Expectation Maximization

Various inference approximation techniques may be applied to compute the approximate value for the Q function. To avoid local minimum problem, we have preferred EM on ICM. By considering this distribution as Gaussian,

$$\log p(d_i|d_j, \hat{\tau}_{i,j}) = -\frac{1}{2} \log 2\pi - \log \hat{\tau}_{i,j} - \frac{|d_i - d_j|^2}{2\hat{\tau}_{i,j}^2} \tag{10}$$

$$\log p(z_i|d_j, \hat{\sigma}_i) = -\frac{1}{2} \log 2\pi - \log \hat{\sigma}_i - \frac{|Z_R(i) - Z_L(i + d_i)|^2}{2\hat{\sigma}_i^2} \tag{11}$$

Free energy is a lower bound on the posterior distribution which is same as minimizing $Q(h)$. In order to minimize energy of Q function, replace values of equation Eq.10,11 in Eq.9 and by equalizing the derivative to zero,

$$\frac{\partial F(d_i, d_j)}{\partial Q(d_i = a|d_j = b, \hat{\tau}_{i,j})} = 0 \tag{12}$$

$$Q(d_i|d_j, \hat{\tau}_{i,j}) \propto \exp \frac{|d_i - d_j|^2}{2\hat{\tau}_{i,j}^2} - \frac{|Z_R(i) - Z_L(i + d_i)|^2}{2\hat{\sigma}_i^2} \tag{13}$$

The restriction placed on the Q function is $\int_{d_i} Q(d_i|d_j, \hat{\tau}_{i,j}) = 1$ (all the constant terms which do not participate in the optimization are not taken into consideration). Uncertainties calculated from Q will decide the significance of the pixel. Intuitively the efficiency of σ_i and $\tau_{i,j}$ can be visualized as competing parameters. For any occluded pixel and its matched pixel, their difference of intensity is a high value. Their variance is also high resulting the weight of the pixel as a low value. Likewise, if the difference of disparities of i and j is high, its influence will be controlled by $\tau_{i,j}$. To minimize σ_i and $\tau_{i,j}$, take derivative of the free energy $F(Q, P)$ and place it equal to zero. Minimum values for the parameters can be found by taking derivative of free energy with respect to them and equating it with zero.

$$\begin{aligned} \hat{\tau}_{i,j}^2 &= \int_{d_j} Q(d_j) \int_{d_i} Q(d_i|d_j, \hat{\tau}_{i,j}) |d_i - d_j|^2 \\ \hat{\sigma}_i^2 &= \int_{d_i} Q(d_i) |Z_t(i) - Z_{t-1}(i + d_i)|^2 \end{aligned} \tag{14}$$

While the constraint placed on $Q(d_i)$ is defined as:

$$Q(d_i) = \int_{d_j} Q(d_j) Q(d_i|d_j, \hat{\tau}_{i,j}, \hat{\sigma}_i) .$$

2.5 Dynamic Programming on Tree

Energy for Q function at node d_i can be expressed as:

$$EQ(d_i|d_j, \hat{\tau}_{i,j}) = Q(d_i|d_j, \hat{\tau}_{i,j}) \prod_{k \in C} Q(d_k|d_j, \hat{\tau}_{k,j}) \quad (15)$$

Minimum energy can be achieved by minimizing log of above expression. Optimal disparity assignment for node can be determined by using:

$$EQ(d_i|d_j, \hat{\tau}_{i,j}) = \arg \min_{d_i \in D} \left\{ s(d_i, d_j) + m(d_i) \sum_{k \in C_i} EQ(d_k|d_j, \hat{\tau}_{k,j}) \right\} \quad (16)$$

Where $s(d_i, d_j)$ is the disparity mismatch and $m(d_i)$ is the matching penalty for assigning disparity d_i to pixel i . Eq. 16 is a standard expression for the optimization [14] problem and can find the optimal places for the minimum energies at each node. Total computation cost for these terms is $O(l^2n)$ each where n is number of nodes and l is max possible disparity vector. Including free energy, total computation cost is $(2m + 1)O(l^2n)$ where m is number of iteration. While implementing this algorithm, the hardest problem is memory consumption. Finding disparity value for each pixel for l different places is hard on memory.



Fig. 2. Comparison Results: Occlusion results for "tsukuba"

3 Experiment Results and Discussion

To check the performance in modeling and detection of occlusion, we test our results against Middlebury test bed [11] dataset. We tested Tsukuba, Sawtooth, Venus and Map image pairs and their results are shown in Fig 2. As compared to other approaches, our results lay in range of position 8 to 10 out of 36 competitors for various images. Although the ranking looks odd, but it's a bit unfair to compare our results straight with other state of art algorithms. Since they employed 2D optimization, while in our case, we are trying to achieve the efficiency of 2D using a tree structure which is neither 2D nor 1D. Its straight comparison can be made either with 1D optimization algorithms or with Dynamic Tree Optimization algorithm [14]. There are 4 methods based on 1D optimization in the evaluation table, and by a coincidence they have consecutive ranks 25 to 28, which is almost at the 3rd quarter of the table. Tree Optimization method

[14] lies a bit high of these algorithms. Further we have incorporated occluded pixels in our results while other results are based on non occluded pixels present in the image pairs. These methods isolate all occluded pixels by using a fix threshold or a threshold based on the neighboring pixels. To look into deep , we tried to find equal footings to compare our occlusion results with several recent approaches: "GC+occl" algorithm by Kolmogorov and Zabih [10] which is a pixel-based approach using a symmetric graph-cut framework to handle occlusion, "Seg+GC" algorithm by Hong and Chen [12] which is a segment-based asymmetric graph-cut approach that does not explicitly detect occlusion, and "Layer" algorithm by Lin and Tomasi [4] which is a combination of pixel-based and segment-based approaches. Results of two images i.e. "tsukuba" and "venus" are presented from[11] dataset. Same parameters are selected for both data sets. The occlusion result is computed by check-ing the visibility of each point in the non-occlusion result. Result of "Layer" is from the authors' website. The results are shown and compared in Fig.3. 1 gives the error statistics for "tsukuba" and "venus" respectively. They are quantitatively evaluated by 3 criteria, which are the percentages of: false positive, false negative and the bad points near the occlusion. A bad point is a point whose absolute disparity error is greater than one [11]. We make a near occlusion model by dilating the occlusion area to 20 pixels and excluding the occlusion area. Figure 3 are our results.

Table 1. Error Statistics of Two images with respect to our technique

Methods	False Pos	False Neg	Near Occl
Tsukuba			
Our Results	2.21	31	8.75
GC+Occl	1.49	31.8	6.34
Seg+GC	1.23	30.4	8.12
Layered	2.25	24.2	9.01
Venus			
Our Results	1.2	21	7
GC+Occl	1.91	32.88	13.12
Seg+GC	0.51	16	0.89
Layered	0.32	51	1.01

4 Conclusion

In this work, we have extracted mst based on the color information. We have treated this mst as a bayesian network and inferred two random parameters for occlusion modeling and smoothness in disparity space. For inference, Helmholtz free energies equations are reshaped to suit our framework i.e. the equations are transformed in a recursive manner to fit in for DP. EM algorithm is used to approximate these energies, and then these are minimized to find the optimal disparity using DP. The ultimate goal achieved is minimum occlusion.

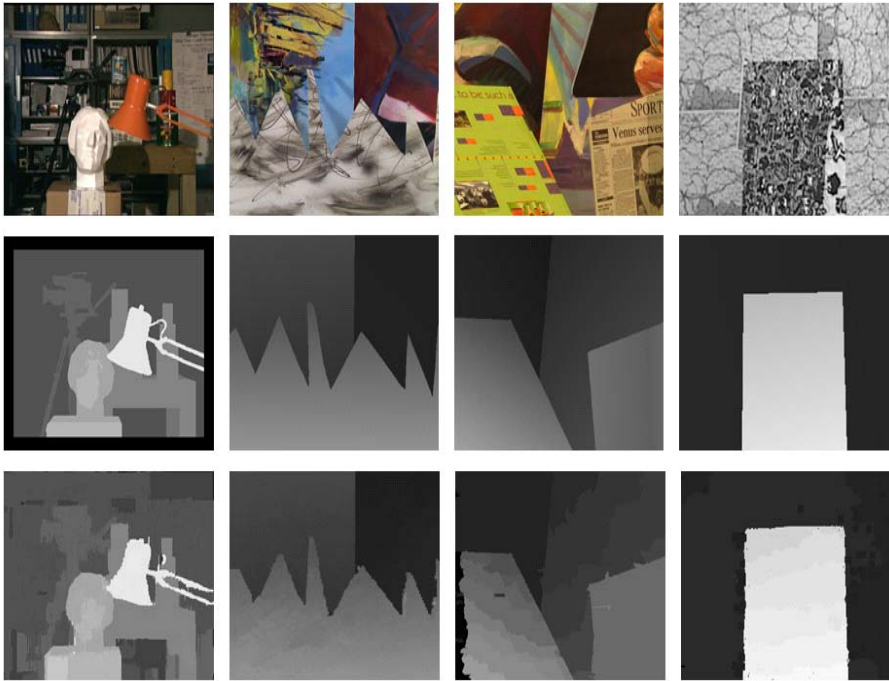


Fig. 3. Comparison Results: Middlebury datasets. First row are the left images, second row are the ground truth third row is our results.

Acknowledgement. This work is jointly sponsored by Higher Education Commission of Pakistan, under National University of Science and Technology, Pakistan.

References

1. P.N. Belhumeur. A bayesian-approach to binocular stereopsis. *IJCV*, 19:237–260, 1996.
2. A.F. Bobick and S.S. Intille. Large occlusion stereo. *IJCV*, 33(3):1–20, 1999.
3. Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *TPAMI*, 11(23):1222–1239, 2001.
4. Myron Z. Brown, Darius Burschka, and Gregory D. Hager. Advances in computational stereo. *TPAMI*, 7, 2003.
5. T. Cormen, C. Leiserson, and R. Rivest. *Introduction to Algorithms*. MIT Press, 1990.
6. I.J Cox, S.L. Hingorani, S.B Rao, and B.M. Maggs. A max likelihood stereo algorithm. *Computer Vision and Image Understanding*, 63:542–567, 1996.
7. Brendan J. Frey and Nebojsa Jojic. A comparison of algorithms for inference and learning in probabilistic graphical models. *TPAMI*, 27(9), 2005.
8. D. Geiger, B. Ladendorf, and A. Yuille. Occlusions and binocular stereo. *IJCV*, 14(3):211–226, 1995.

9. D.Geiger H.Ishikawa. Occlusions,discontinuities,and epipolar lines in stereo. In *ECCV*, 98.
10. V. Kolmogorov and R. Zabih. Multi-camera scene reconstruction via graph cuts. In *ECCV*, 02.
11. D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 1-3(47):7-42, April 2002.
12. Milan Sonka, Vaclav Hlavac, and Roger Boyle. *Image Processing Analysis and Machine Vision*. USA, 2nd edition, 2002.
13. J. Sun, N. Zheng, and H. Shum. Stereo matching using belief propagation. *TPAMI*, 7, 2003.
14. Olga Veksler. Stereo correspondence by dynamic programming on a tree. In *CVPR*, 2005.

Estimation of Rotation Parameters from Blurred Image

Qian Li and Shi-gang Wang

Mechatronics Design & Automation Technology Institute, School of Mechanical Engineering, Shanghai Jiao Tong University, 200030, P.R. China
Liqian@sjtu.org, Wangshigang@sjtu.edu.cn

Abstract. Many industrial applications involve rotations. Different from traditional measurements, we propose a novel vision method based on image motion blur to estimate angular velocity and angular displacement in this paper. First, we transform 2D rotation to 1D translational motion by sectoring rotation blurred image. Then we use mathematical models in spatial and frequency domain to analyze translation blurred images. According to mathematical models in frequency domain, there is a series of dark parallel lines on the spectrum of the translation blurred image. These dark lines are exactly related to the velocity of translation and exposure time. Furthermore, based on the geometric relationship between rotation and translation, these dark lines are also related to angular velocity. Both simulation results and real experimental results, based on the proposed method in this paper, are provided. These results demonstrate the feasibility and efficiency of proposed method.

1 Introduction

Investigation of a rotating machinery and the equivalent rotational vibration of systems are very important for control systems in industry [1]. Traditional angular and rotary velocity sensors are contact-type tachometers or speed sensors that mount on a shaft or contact a moving surface, such as mechanical tachometers, electrical tachometers [2,3], etc. Because classical methods are usually based on mechanical contact, they are easily affected by the target's motion. Over the past decade, non-contact methods have been developed for industrial applications [4]. The non-contact sensors available today use circular Moire gratings [5], tomography [4], magnetic method, ultrasound, laser, etc. Although non-contact sensors overcome these disadvantages of conventional contact measurements, most of them require particular additional equipments, such as laser sources and reflectors. For measurement tasks, these additions may make a measuring system very complicated and expensive.

In recent years, some researchers have begun to investigate angular velocity estimation based on digital image processing techniques. Yamaguchi [6,7] proposed a gaze control active vision system integrated with angular velocity sensors to extract the velocity. The velocity is obtained from a continuous series of images acquired by a visual system.

For real image system, due to the relative motion between a camera and an object within finite exposure time, such images are degraded by well-known degradation factor called motion blur. In previous researches, identification of blur parameters from motion blurred images was used to restore image [8,9,10,11,12,13]. It makes sense to note that certain motion information can be involved in a motion blurred /rotation blurred image. It is natural to use this "blur" knowledge to estimate rotation parameters from blurred images recorded by the camera. Although motion blur is generally considered as extra source of noise and most researchers try to avoid it in conventional motion estimated method, some re-searchers estimated the motion using blurred images. Chen [14] established a computational model in the frequency domain to estimate image motion from motion blur information, which involved a special sensor system to avoid zeros of Point Spread Function (PSF) in the frequency domain.

In this paper, we investigate the important visual information— translation and rotation blur—for the parameter estimation of rotation and propose a novel non-contact method of angular velocity and the angular displacement estimation with singular blurred image. Compared with Chen [14], our method does not need to avoid zeros of PSF; on the contrary, we utilize these zeros to achieve the purpose. In order to reveal the relationship between rotation and image blur, we give the geometric relations between the rotation and the translation, and then transform the 2D rotation to a 1D translational motion by sectoring the rotation blurred image. After establishing mathematical models in the spatial and frequency domain of translation blurred images, we found there is a series of dark parallel lines on the spectrum of it, which was sectored from the rotation blurred image. According to the geometric relationship, these dark lines are exactly related to the angular velocity of rotation. By extracting the lines' information with image processing techniques, we estimated the angular velocity.

2 Rotation Blurred Image Analysis

Estimation of rotation parameters from a blurred image are extracted in three steps. First, we transform rotations to translations. Second, the mathematic model of translations in the spatial and frequency domain is established. Next, the direct relation between blur information and motion parameters is extracted.

2.1 Transform Rotation to Translation

We introduce a polar coordinate system (Fig. 1(a)) to represent the image plane. The polar coordinate of image point is (r, θ) . Then we transform the polar lattice of image plane to a rectangular lattice (x, y) (Fig. 1(b)) by introducing radial pixels, where θ is the angular resolution and r is the resolution along the radius. Let us introduce the polar system whose origin O lies on the rotary center of target object. The largest radial pixel determines the sampling step of a rectangular system. According to Fig. 1, these positions of pixels in polar and rectangular lattices do not match. So we used bilinear interpolation to determine the

gray-level of radial pixels. The gray level for a rectangular pixel $f(i, j)$ is calculated from the gray level of its four neighboring polar pixels as:

$$f(i, j) = (1 - t)(1 - u)f(x, y) + t(1 - u)f(x + 1, y) + (1 - t)uf(x, y + 1) + tuf(x + 1, y + 1) \tag{1}$$

where:

$$x = \text{floor}(i) \quad t = i - x \tag{2}$$

$$y = \text{floor}(j) \quad u = j - y \tag{3}$$

As well known bilinear interpolation (zoom) is one of the most common algorithms used in image processing. It is a fractional zoom calculation and used to change the size of an image while retaining good image quality.

So, to ease the estimation of rotary motion based on the blurred image, we transform the rotary motion blurred image to polar lattice. Thus we sector it at the rotary center. After sectoring, the image on polar coordinates is opened

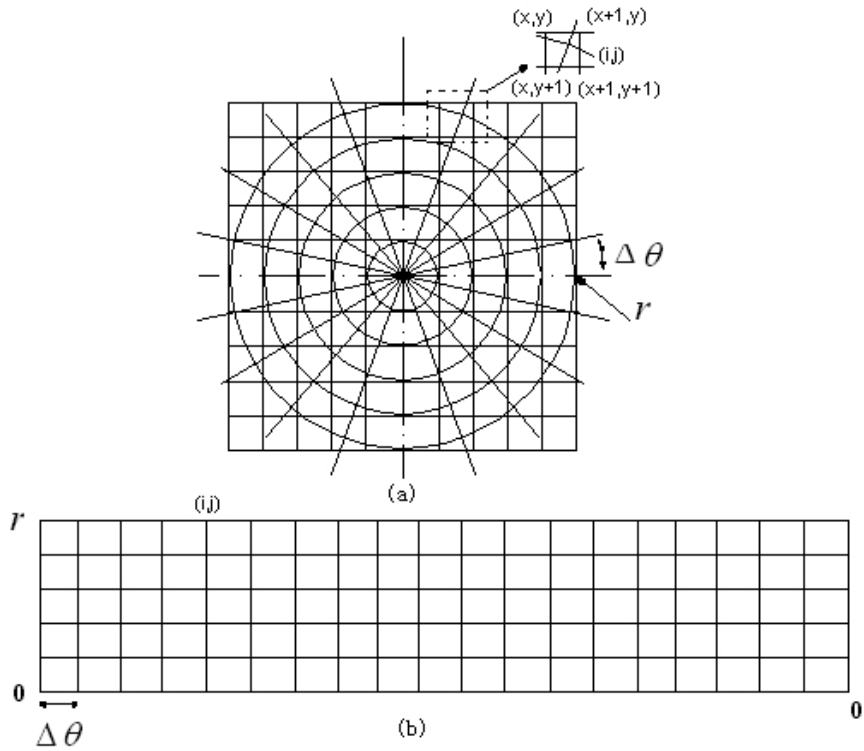


Fig. 1. Polar-to-rectangular lattice transformation. (a) Polar lattice system. (b) Rectangular lattice system.

as a sector. For rotary motion, the angular displacement related to the static position is approximately expressed as:

$$\theta_m(t) = \varphi_0 + \omega(t) \tag{4}$$

where ω is the angular velocity, φ_0 is the angular displacement at the beginning of exposure time from the static position.

Assuming r is the rotary radius, the linear velocity is expressed as:

$$v = r\omega(t) \tag{5}$$

here, points on the blurred image are defined by velocity $(v_{i,j})$.

So the movement of points on the moving object with the same rotary radius is considered as a linear uniform motion. If we sector the motion blurred image along the normal direction by the method we discussed before, the obtained image is the translation blurred image. The 2D problem is thus reduced to a 1D problem. The velocity is measured by detecting the linear velocity of obtained image. In the next part, we will discuss how to estimate the translation based on the motion blurred image.

2.2 The Fundamental Assumption

An image is formed when an image detector integrates light energy over a time interval [8]. If illumination is stable and target objects are diffused and reflective surface, we assume that the light intensity I_s projected from the same physical point of object to the CCD image plane will not change whether it is in motion or not. The beginning exposure time is defined as the starting time, and the position (x, y) of a point at that time is defined as the initial static position. In case of arbitrary motion during exposure time t , displacements related to the initial static position are expressed as $(x_0(t), y_0(t))$, along the x-direction and y-direction, respectively. When the point moves from (x, y) to $(x_0(t), y_0(t))$ during the exposure time t , the light intensity can be expressed as:

$$I(x + x_o(t), y + y_0(t), t) = I(x, y, t_0) \tag{6}$$

By variables transforming (x, y) , Equation(6) can be expressed as follows:

$$I(x, y, t) = I(x - x_o(t), y - y_0(t), t_0) \tag{7}$$

Defining $I(x, y, t_0)$ as the light intensity $I_s(x, y)$ of static image, we have

$$I(x, y, t) = I_s(x - x_o(t), y - y_0(t)) \tag{8}$$

2.3 The Motion Blurred Image in Spatial Domain

The gray level of any point on the image recorded by CCD is proportional to the integral of light intensity at this point during the exposure time. So the gray level of static image is:

$$f(x, y) = k \int_0^{t_e} I_s(x, y) dt = kI_s(x, y)t_e \tag{9}$$

where k is the photoelectric transformation coefficient of CCD, and t_e is the exposure time.

According to the linear integral character of CCD, the motion blurred image $g(x, y)$ can be regarded as the superposition of motion object at every instant position on an image plane. Within an interval in exposure time $[t_i, t_i + dt]$, the motion object is considered as quiescent condition. So the small increment of gray level is $dg(x, y)$:

$$dg(x, y) = kI(x, y, t)dt = kI_s(x - x_0, y - y_0(t))dt \quad (10)$$

Substituting (9) for (10) yields

$$dg(x, y) = \frac{1}{t_e} f(x - x_0(t), y - y_0(t))dt \quad (11)$$

By integrating two sides of (11) during the exposure time, the relationship between the static image and its blurred version can be expressed as:

$$g(x, y) = \frac{1}{t_e} \int_0^{t_e} f(x - x_0(t), y - y_0(t))dt \quad (12)$$

2.4 Translation of Motion Blurred Image in Frequency Domain

Assuming $x_t(t)$ and $y_t(t)$ as translation components along x and y direction, from (12), the relationship between the linear uniform motion image and the static image in the spatial domain is:

$$g(x, y) = \frac{1}{t_e} \int_0^{t_e} f[x - x_t(t), y - y_t(t)]dt \quad (13)$$

After the Fourier transform:

$$G(u, v) = \frac{1}{t_e} F(u, v) \int_0^{t_e} e^{[-i2\pi(ux_t(t)+vy_t(t))]} dt \quad (14)$$

We define $H(u, v)$ as:

$$H(u, v) = \frac{1}{t_e} \int_0^{t_e} e^{[-i2\pi(ux_t(t)+vy_t(t))]} dt \quad (15)$$

Then, Equation (14) can be simplified to (16):

$$G(u, v) = F(u, v)H(u, v) \quad (16)$$

If there is motion just in the x direction, the velocity is $v_x(t) = a/t_e, v_y = 0$ then the displacement is $x_t(t) = at/t_e, y_t(t) = 0$. When $t = t_e$, the displacement of image is a . So Equation (15) becomes:

$$H(u) = \frac{1}{t_e} \int_0^{t_e} e^{(-i2\pi uat/t_e)} dt = \frac{1}{\pi ua} \sin(\pi ua) e^{(-i\pi ua)} \quad (17)$$

$H(u)$ is the point spread function (PSF) of motion. The zeroes of PSF can be written as:

$$\mu_i = \pi ua \tag{18}$$

These standard values of zeroes μ_i can be known by their sinusoidal function, that is $n_i = ua, n_i = 1, 2, \dots, n$. Based on this analysis, there is a series of darkness parallel lines on the spectrum that relate to these zeroes of $H(u)$. These lines are extracted by Radon transformation of image processing in this paper. Meanwhile, these lines are vertical in the motion direction. So the displacement is calculated by $H(u)$'s zeroes. The mathematical expression shows that increasing the displacement of motion creates more lines in the frequency domain, and decreasing the displacement of motion decreases the number of these lines. Based on the geometric relationship between translation and rotation, we obtained relation between the displacement of translation and the angular velocity, and the exposure time.

$$\omega(t) = a/rt_e \tag{19}$$

Where, r is the radius of geometric transformation. The angular velocity can be obtained by (19).

2.5 The Method for Simulating Blurred Image

For the translational motion with $v_x(t)$ and $v_y(t)$ as motion components along x and y direction, from the (13), we get:

$$\begin{aligned}
 g(x, y) &= \frac{1}{t_e} \int_0^{t_e} f(x - x_t(t), y - y_t(t)) dt \\
 &\cong \frac{1}{t_e n \Delta T} \sum_{i=0}^{(n-1)\Delta T} f(x - v_x i \Delta T, y - v_y i \Delta T)
 \end{aligned}
 \tag{20}$$

Where the time interval $[0, t_e]$ is divided into n steps, and every time step is ΔT . So the translation blurred image $g(x, y)$ is simulated by summing up unblurred images $f(x - x_t(t), y - y_t(t))$.

For the rotation blurred image, we represent the image plan by polar coordinates system. The polar coordinate of image point is (r, θ) , where the polar center is located in the center of blurred image, and $r = (x^2 + y^2)^{1/2}$, (x, y) is the Cartesian coordinate of image point. Then, let the rotational center of blurred image locate at the point of polar center. The simulated image of rotation blur is deduced using the same procedure in (20).

$$\begin{aligned}
 g(r, \theta) &= \frac{1}{t_e} \int_0^{t_e} f(r, \theta - \theta(t)) dt \\
 &\cong \frac{1}{t_e n \Delta T} \sum_{i=0}^{(n-1)\Delta T} f(r, \theta - i\omega \Delta T)
 \end{aligned}
 \tag{21}$$

Where the time interval $[0, t_e]$ is divided into n steps, and every time step is ΔT . So summing up unblurred images simulates the rotation blurred image $f(r, \theta)$.

3 Experimental Results

We evaluate the method's ability to estimate the angular velocity and the angular displacement from the blurred image with simulated images and real images.

3.1 Simulation

In this section, two simulated experiments are carried out to verify the proposed estimation of translation displacement and the angular displacement, using the procedure in (20) and (21).

In first experiment, based on the procedure in (20), the simulated exposure time t_e is 1s, and then we let $v_x\Delta T = 1 \text{ pixel}$, and $v_y\Delta T = 0$. Binary English character images at a resolution of 165×165 pixels and at a resolution of 320×320 pixels, as shown in Table 1., are used. Meanwhile, Table 1. shows estimation results for English characters with $n = 25$, $n = 30$, $n = 40$, respectively. That is, displacements of them are 25 pixels , 30 pixels , 40 pixels and velocities are 25 pixel/s , 30 pixel/s and 40 pixel/s respectively.


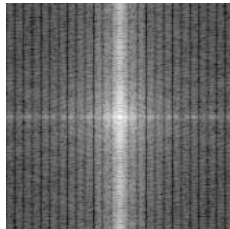

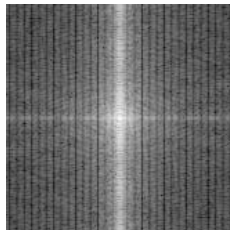

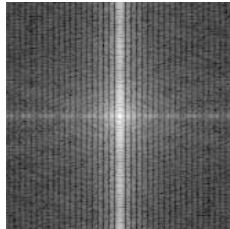
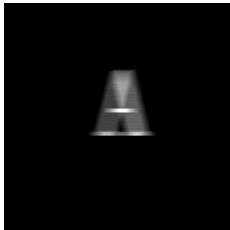
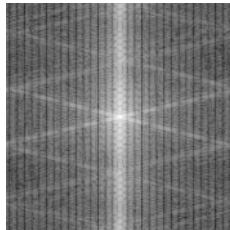
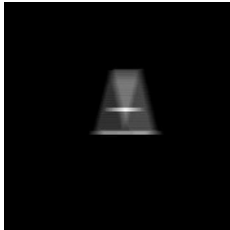
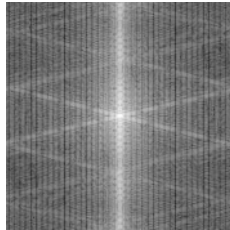
These results of the first experiment are given in Table 1. From Table 1, we found that the maximum angular relative error is 0.093. When we increased the resolution of blurred image, the relative error decreased to 0.004. This proves that our algorithm has good accuracy.

In the second simulated experiments, an IC element image at a resolution of 320×320 pixels in electronic manufacture is applied to obtain a rotation blurred image. In this experiment, the simulated exposure time t_e is 1s, and $\omega\Delta T = 1^\circ$. Table 2 shows estimation results for IC image with $n = 20$. The angular displacement is 20° and the angular velocity is 0.349 rad/s . Figure 2 shows the simulated rotation blurred image and sectored image of it. Figure 3 shows the spectrum of sectored image.

The steps to estimate the angular displacement are as follows:

1. Rotation center detection. We extracted the rotation center by traditional circles detection algorithms. Two-steps algorithm has been used to detect the circle from pairs of intersecting chords using Hough transform (HT), which is robust against object defects and shape distortions. In the first step, we applied the Sobel edge detection algorithm to find circles' edges in the rotation blurred image. Then, the 2D HT is used to compute the centers of the circles. The points computed from chords are being voted to the 2D accumulator array, the significant peak is detected as the center.
2. Sectoring the rotation blurred image. We sector the rotation blurred image along the normal direction at the rotary center based on the algorithm in Sect 2.1.
3. Lines Detection. We extract these parallel lines on spectrum using the Radon transformation. The Radon transform is a standard tool to extract parameterized straight lines from images.
4. Calculating results. After positions of parallel lines are found, the angular displacement or angular velocity is obtained by (18) and (19).

Table 1. Experimental results of translation velocity estimation

Blurred Image	Sectored Image	Velocity	Results	Relative Error
		25pixel/s	27.333pixel/s	0.093
		30pixel/s	32.8pixel/s	0.093
		40pixel/s	41pixel/s	0.025
		30pixel/s	30.12pixel/s	0.004
		40pixel/s	40.02pixel/s	0.0005

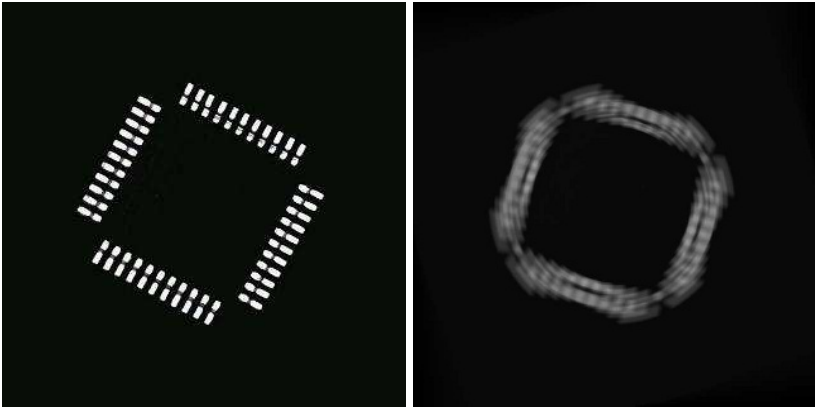


Fig. 2. Original IC element Image and Simulated rotation blurred image with displacement of 20 degree

Table 2. Experimental results of translation velocity estimation

Lines No.	1	2	3	4
Angular displacement ($^{\circ}$)	19.15	18.80	19.15	19.15
Lines No.	5	6	7	8
Angular displacement ($^{\circ}$)	19.15	19.15	18.80	19.15

Results of second experiment are given in Table 2. These results of angular displacement are calculated by the position information, which is detected by image processing methods. The average of these calculated results is 19.06° and the relative error is 0.042. And the angular velocity is $0.33rad/s$ and the relative error is 0.048. These results show that our method is quite accurate when applied to rotation.

3.2 Real Experiment

In this section, we demonstrate the feasibility of the novel method for a real-world image taken by a camera.

The schematics for this rotation estimation and part of real experiment equipment are shown in Fig. 4.

An axle head of servomotor is applied to generate a blurred image. The exposure time of camera is controllable by software, so here, we set the exposure time as 0.01618s. First, we keep the servomotor rotated, the revolution number of it

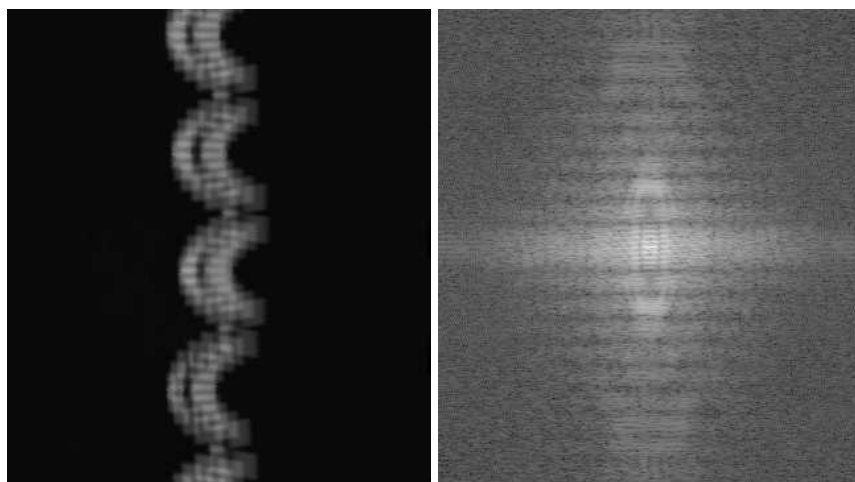


Fig. 3. Sectored blurred image of simulated rotation-blur image and Spectrum of sectored blurred image

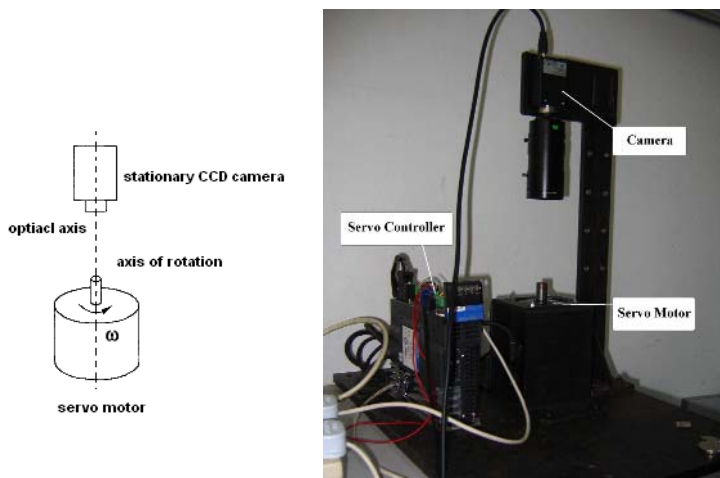


Fig. 4. The schematics for rotation estimation and part of real experiment equipment

is $400r/min$ that is correspond to the feedback value measured by photoelectric encoder in the servomotor. And then the camera acquires the blurred image. The blurred image at a resolution of 479×479 , the sectored blurred image and its spectrum are all shown in Fig.5.

These results of revolution number are calculated by the position information, which is detected by image processing methods. The average of these calculated results is $387.06r/min$ and the relative error between calculated value obtained from the vision method in this paper and the measured value obtained from

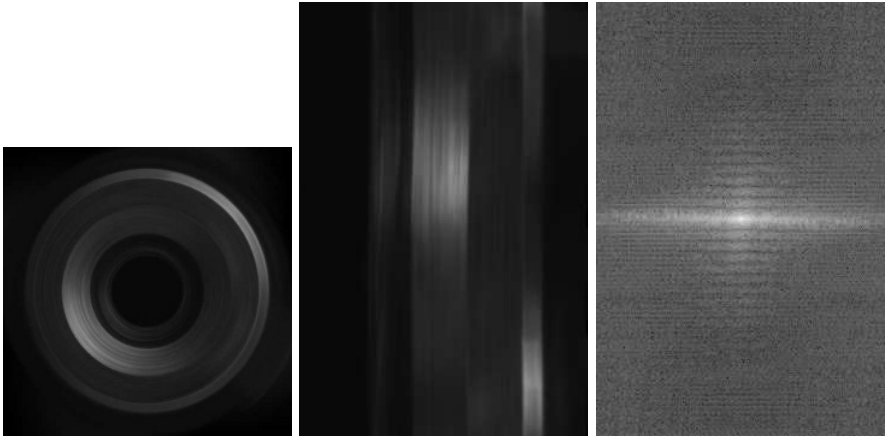


Fig. 5. Real experiment revolution number is $400r/min$

photoelectric encoder is 0.032. These results show that our method is efficiency when applied to rotation.

4 Conclusions

We have presented a new method for rotation analysis from singular blurred image using “motion blur” information. We utilize the motion blur of the motion image rather than consider it as degradation on image. This method calculated the angular velocity and angular displacement from singular blur image. Further, our algorithm decomposes a 2D rotation analysis problem into a 1D translation problem, which simplifies the rotation analysis. The agreement between proposed method and experimental results demonstrates the feasibility and efficiency of this method.

Acknowledgement. It is gratefully acknowledged that the National Natural Science Foundation of China provided funds (Grant No.: NSFC 50375099 & NSFC 50390064) for the financial support of this work.

References

1. Marc Bodson, John Chiasson, and Robert T. Novotnak: Nonlinear speed observer for high-performance induction motor control. *IEEE Transactions on industrial electronics*, **42**(4),(1995): 337-343
2. Polak, T.A.; Pande, C. *Engineering Measurements - Methods and Intrinsic Errors*, John Wiley & Sons, (1999)
3. Ammar Hadi Kadhim, T.K.M. Babu, and Denis O’Kelly, Measurement of steady-state and transient load-angle, angular velocity, and acceleration using an optical encoder, *IEEE Transactions on instrumentation and measurement*, **41**(4)(1992): 486-489

4. Emmanuel O.Etuke and Roger T.Bonnecaze, Measurement of angular velocities using electrical impedance tomography, Flow measure and instrumentation, **9(3)**(1998):159-169
5. Yun Long Lay, Wen Yuan Chen, Rotation measurement using a circular moiré grating, Optics & Laser Technology, **30**(1998): 539-544
6. Teruo Yamaguchi, Hiro Yamasaki, Active vision system integrated with angular velocity sensors, Measurement, **15**(1995): 59-68
7. Teruo Yamaguchi, Hiro Yamasakit, Velocity based vestibular-visual integration in active sensing system, Proceedings of the 1994 IEEE International conference on multisensor and integration for intelligent systems, Oct.2-5 (1994): 630-637
8. Moshe Ben-Ezra and Shree K. Nayar, Motion-based motion deblurring, IEEE transactions on pattern analysis and machine intelligence, **26(6)**(2004): 689-698
9. Michal Haindl, Recursive Model-Based Image Restoration, Pattern Recognition, 2000. Proceedings. 15th International Conference on (2000). 342 - 345.
10. Yagnesh C.Trivedi and Ludwik Kurz, Image restoration using recursive estimators, IEEE Transactions on systems, man, and cybernetics, **25(11)**(1995):1470-1482
11. Moon Gi Kang, Adaptive Iterative Image Restoration Algorithms, Diss. 1994
12. A.K.katsaggelos, Iterative image restoration algorithm, Optical engineering **28(7)**(1989):735-748
13. Katsaggelos, A., Biemond, J., Mersereau, R. and Schafer, R., A general formulation of constrained iterative restoration algorithms Acoustics, Speech, and Signal Processing, IEEE International Conference on ICASSP 85. **10**(1985): 700 - 703
14. Wei-Ge Chen, N. Nadhakumar and Worthy N. Martin, Image motion estimation from motion smear- a new computational model, IEEE transactions on pattern analysis and machine intelligence, **18(4)**(1996): 412-425

Hierarchical Stereo Matching: From Foreground to Background

Zhang Kai, Wang Yuzhou, and Wang Guoping

HCI & Multimedia Lab., School of Electronics Engineering and Computer Science,
Peking University, 100871, China
zk@graphics.pku.edu.cn

Abstract. In this paper we propose a new segment-based stereo matching algorithm using scene hierarchical structure. In particular, we highlight a previously overlooked geometric fact: the most foreground objects can be easily detected by intensity-based cost function and the farer objects can be matched using local occlusion model constructed by former recognized objects. Then the scene structure is achieved from foreground to background. Two occlusion relations are proposed to establish occlusion model and to update cost function. Image segmentation technique is adopted to increase algorithm efficiency and to decrease discontinuity of disparity map. Experiments demonstrate that the performance of our algorithm is among the state of the art stereo algorithms on various data sets.

1 Introduction

Stereo matching is one of the most active research areas in computer vision. It serves as an important step in a wide range of applications such as view synthesis, image based rendering and modeling and robot research, etc. Given a pair of horizontal registered images from the same scene, the goal of stereo matching is to determine the dense disparity map. For every 3D point P in the scene, if it is not occluded in both images, we can find a pixel p in the left image which corresponds to P , and a pixel q in the right image too. The two pixels lie on the same horizontal scanline since the images are rectified. The difference in the horizontal position of p and q is termed as disparity, which is in inverse ratio with P 's depth [1,2]. Disparity map consists of all pixels' disparities.

1.1 Previous Work

There exists a considerable body of work on the dense stereo correspondence problem. [1,2] have provided an exhaustive comparison of dense stereo correspondence algorithms. In general, stereo algorithms can be categorized into two major classes: the methods based on local constraints and the methods based on global constraints. Block matching methods seek to estimate disparity at a point in one image by comparing a small window about that point with a series

of small windows extracted from the other image. Different methods use different windows, like solid windows [3], multiple windows [4,5], adaptive windows[6]. Local methods can be very efficient, but they are sensitive to locally ambiguous regions in images (e.g., textureless regions, disparity discontinuous boundaries and occluded portions).

Global methods can be less sensitive to ambiguous regions since global constraints provide additional support for regions difficult to match locally. Dynamic programming is a mathematical method that reduces the computational complexity of optimization problems by analyzing the corresponding scanlines and constructing appropriate occlusion model [7,8,9]. The most significant limitation of dynamic programming for stereo matching is its inability to strongly incorporate both horizontal and vertical continuity constraints. An alternative approach that exploits these constraints is to cast the stereo matching problem as that of finding the maximum flow [10] in a graph [11,12,13,14]. Graph cut methods have been shown to be among the best performers in [1]. However, these methods are more computationally expensive because of too many iterations.

A lot of segment-based stereo algorithms arise recently [15,16,17], which are based on the assumption that there are no large disparity discontinuities inside homogeneous color segments. Image segmentation representation is used to reduce the high solution space and enforce disparity smoothness in homogeneous color regions. Usually the segmentation technique is integrated within other frameworks and achieves strong performance. The approach we propose is also segment-based.

1.2 Overview of Our Approach

The proposed algorithm is inspired upon two facts: dynamic programming and image segmentation. Dynamic programming is well known for its efficiency and it can take advantage of all knowledge of the occlusion geometry in one scanline to model occlusion relation for later correspondence. Image segments contain much richer information than one single scanline, and the possibility of making a wrong decision upon a segment could be greatly reduced. The two facts motivate us to model an accurate occlusion relation based on image segment then to offer a better algorithm.

In our approach, the scene structure is divided into a set of disparity levels. The stereo matching problem becomes assigning the corresponding disparity to each level, which can be easily formalized as an cost minimization problem in the segment domain. Specifically, the cost function contains two levels: pixel matching cost eliminates sampling errors, and segment matching cost measures the disagreement of segments.

In the hierarchical scene structure, the nearest objects can be first matched because they are never occluded, then an accurate local occlusion model is constructed by the detected foreground for the farer objects' matching process. Once one level is identified, the pixels in this level are termed as ground control points(GCPs) [8,9,16]. When we finished matching process from foreground to background, the dense disparity map is obtained and invalid disparities are

eliminated. For scenes composed of several planes we also recognize each plane by segment coalition method and plane fitting equation. It is worth to notice that while the algorithm computes disparity map for each image using consistency constraint, the generalization for multiple input images is straightforward.

The rest of the paper is organized as follows: First some stereo constraints (section 2) and image segmentation technique (section 3) are briefly described. Then the proposed algorithm is presented in detail mainly focus on how to define cost function (section 4) and how to apply occlusion relation to assign the corresponding disparity to each scene disparity level (section 5). We provide various experimental results in Section 6 to demonstrate the algorithm's strong performance for traditional challenging image regions. Finally, we conclude in Section 7.

2 The Definitions of Constraints

Here we introduce three consistency constraints. These constraints serve two purposes: they facilitate a reasonable search of possible match, and they disallow certain types of unlikely match. First of all, several concepts in stereo vision are re-formulated and generalized to facilitate the definitions of the constraints. The definitions of constraints are given later.

We use a pair of horizontally rectified stereo images to ease the description of the algorithm through out the paper. Let I_l denote the left image and I_r the right image. Let D_l denote the disparity map of I_l and D_r of I_r . D_l is a function that assigns each pixel p in I_l a disparity d , a horizontal displacement vector: $D_l(p) = d$, such that d is the disparity of $q = p + d$ in I_r : $D_r(p - d) = d$. Let $I_l(p)$ denote the intensity of the left image pixel p and $I_r(q)$ the intensity of the right image pixel q . At the beginning, D_l and D_r are both empty. A pair (p, d) is termed as a match, which could be considered as a 3D point.

Definition 1 (Uniqueness Constraint). *Every pixel has one and only one disparity: $D_l(p) = d_1 \wedge D_l(p) = d_2 \implies d_1 = d_2$.*

Definition 2 (Ordering Constraint). *The ordering of two pixels of a scanline in one image is kept in the other image: $p_l < p_r \implies p_l - D_l(p_l) < p_r - D_l(p_r)$.*

Definition 3 (Consistency Constraint). *Corresponding pixels share the same disparity: $D_l(p) = d \implies D_r(p - d) = d$.*

It can be shown that the above constraints hold for most of the previously used stereo datasets [1,3,7,8]. A detailed discussion is outside the scope of this paper.

3 Segmentation

Our approach is built upon the assumption that large disparity discontinuities only occur on the boundaries of homogeneous color segments. Therefore any

image segmentation algorithm that decomposes an image into homogeneous image segments will work for us. In our current implementation, mean-shift image segmentation algorithm [18] is used.

We assume that pixels inside the same segment have the same disparity and our algorithm actually assigns each segment a disparity. This assumption makes our method very simple and efficient. The assumption seems quite restrictive since it is only valid for fronto-parallel surfaces and becomes problematic when a segment represents a pronounced slanted surface or crosses surface boundaries. However, we claim that, the limitation could be significantly alleviated and the assumption becomes a very good approximation in practice by taking over-segmentation. In fact in slant plane cases we will incorporate all segments on the same plane to improve smoothness as discussed in subsection 5.5. Fig.1 shows segment result of standard dataset Tsukuba [1].

Except for image segmentation, we take an additional segment splitting method that assigns individual pixel special disparity and splits it from the segment dynamically. The splitting method helps capturing object boundaries missed by color segmentation and decomposes a pronounced slanted surface into small regions, therefore making the constant disparity assumption a good approximation and making our algorithm more flexible. Such method is described in subsection 5.3.



Fig. 1. The Tsukuba image. Left column is the reference view. Middle column is the segment result. Right column is initial disparity map computed by segment cost function.

4 Cost Function

Cost function is used for disparity estimation and is usually based on pixel intensity. The cost function of the proposed algorithm contains two levels: pixel matching cost which eliminates sampling errors and segment matching cost which measures the disagreement of segments.

4.1 Pixel Matching Cost

The simplest cost function is as the absolute difference of intensity between two pixels. Given a possible match (p, d) :

$$Cost^\beta(p, d) = |I_L(p) - I_R(p - d)| \tag{1}$$

However, this measure is inadequate for discrete images, because image sampling can cause this difference to be large wherever the disparity is not an integral number of pixels and in this case every 3D point’s intensity is distributed over several pixels. Typically, the sampling problem is alleviated by using some linearly interpolated intensity functions in a method that are insensitive to sampling (Birchfield and Tomasi, [7]).

In practice, we found Birchfield’s method just alleviated sampling error on one side, but failed when sampling error occurred on both sides especially in textureless regions. We improved their method and achieved better effects. Consider three pixels q_l, q, q_r in right image: $q_l = p - (d + 1), q = p - d, q_r = p - (d - 1)$. We believe that sampling error occurs if and only if the dissimilarity between p and q_l, q, q_r is not greater than a predefined threshold λ :

$$Cost(p, d) = \frac{1}{|\{q_i\}|} \sum_{q_i \in \{q_i\}} Cost^\beta(p, p - q_i) \tag{2}$$

where $\{q_i\} = \{q_i | q_i \in \{p - (d + 1), p - d, p - (d - 1)\} \wedge Cost^\beta(p, p - q_i) \leq \lambda\}$, and $|\{q_i\}|$ denotes the element count of $\{q_i\}$. In our current implement we set $\lambda = 1$.

It is worth pointing out here that our cost function make no use of any type of local window because the segment cost function to be introduced in the next subsection is more effective than any window-based cost function.

4.2 Segment Matching Cost

Segment cost is defined as an average of all pixels’ costs. Consider segment S and let $|S|$ denote the pixel count, so we can write:

$$Cost^\beta(S, d) = \frac{1}{|S|} \sum_{p \in S} Cost(p, d) \tag{3}$$

Just as window-based cost functions, (3) is also a kind of local constraint. However, window-based measurements usually make large errors in the disparity discontinuous boundaries because of occlusion and intensity break [3,4,5,6]. On the contrary in image segment since pixel intensities are all similar, the disparity break is prohibited, so (3) is much more robust than previous window-based cost functions.

When all pixels in S are not occluded, (3) is the standard segment cost function. In subsection 5.2 we give a modified version which concerns the occurrence of occlusion.

5 Hierarchical Stereo Algorithm

In this section we describe the disparity map estimating algorithm in detail. In general our algorithm perform the following four steps. First we obtain the initial disparity map. We then assign the corresponding disparity to each level from foreground to background by modeling proper occlusion relation. Thirdly,

undetermined segments' disparities are identified using GCPs. Finally, if there exist slant planes in the scene we incorporate all segments on the same plane using segment coalition method and plane fitting equation.

5.1 Initialization

Computing the initial disparity map is trivial: compute cost values for all possible matches (S, d) using (3) and simply choose for each segment the disparity associated with the minimum cost and fill in D_l and D_r respectively. Here we assume that all image segments are unoccluded. As shown in Fig.1, our segment cost function gains good initial results.

5.2 Occlusion Relation

In this paper, we emphasize particularly on the occlusion relation for the modification of segment cost function (3). At first, two important occlusion relations in left image are introduced. The corresponding relations in right image can be easily obtained by symmetry.

Definition 4 (Strong Occlusion Relation). *Given two adjoining pixels p_l and p_r of the same scanline in left image, and their corresponding pixels $q_l = p_l - D_l(p_l)$ and $q_r = p_r - D_l(p_r)$ in the right image, if the disparity of p_l is greater than that of p_r , then the pixels between q_l and q_r are all occluded and their disparities are all less than or equal to the p_r 's: $p_l = p_r - 1 \wedge D_l(p_l) > D_l(p_r) \implies \forall q, (q > q_l \wedge q < q_r \implies q \text{ is occluded} \wedge D_r(q) \leq D_l(p_r))$.*

Definition 5 (Weak Occlusion Relation). *Given two adjoining pixels p_l and p_r of the same scanline in left image, and their corresponding pixels $q_l = p_l - D_l(p_l)$ and $q_r = p_r - D_l(p_r)$ in the right image, if the disparity of p_l is greater than that of p_r , then the pixels between q_l and q_r are all occluded and their disparities are all equal to the p_r 's: $p_l = p_r - 1 \wedge D_l(p_l) > D_l(p_r) \implies \forall q, (q > q_l \wedge q < q_r \implies q \text{ is occluded} \wedge D_r(q) = D_l(p_r))$.*

Strong occlusion relation describes the inter-occlusion situation of objects containing more than two kinds of depth, and weak occlusion relation only applies to the case of two kinds of depth. The principle hidden behind both relations is the same: when occlusion occurs, background is occluded by foreground. Or in other words, the larger disparity pixels occlude the smaller ones. Fig.2 shows the two occlusion relations.

In the initialization phase we assume the lackness of occurrence of occlusion. Such assumption would easily bias segment matching cost in occluded portions, but in unoccluded areas it is credible. Obviously the foreground is doubtless unoccluded. Since we represent the scene structure as a set of disparity levels, once a level with larger disparity is recognized, all possible cost values of the segments adjacent to that level should be re-computed. Refer to the two occlusion relations, the occluded pixels in a segment can be decided by GCPs and the given disparity d .

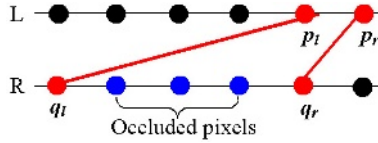


Fig. 2. The occlusion relations. Two horizontal lines are corresponding scanlines of two images. Well-matched pixel(GCP) pairs are in red and linked by red lines. The blue pixels on right scanline is occluded, and their disparities are equal to that of q_r under weak occlusion relation.

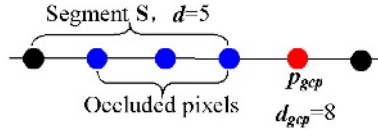


Fig. 3. Occluded pixels. The horizontal line are considered scanline of left image. The red pixel is GCP p_{gcp} with disparity d_{gcp} . the $(d_{gcp} - d)$ blue pixels on the left side of p_{gcp} are occluded.

Definition 6 (Occluded Pixels). To compute cost value of segment S under a given disparity d in left image, for every scanline S strides, we denote the first GCP on the right side of S as p_{gcp} with disparity d_{gcp} . If $(d_{gcp} - d)$ is positive, the $(d_{gcp} - d)$ pixels on the left side of p_{gcp} are occluded. Refer to Fig.3.

It is easy to see that the occluded pixels in segment S is a function of the given disparity d and denoted as $S_{occ}(d)$. Accordingly, the unoccluded pixels are denoted as $S_{unocc}(d)$, and $S = S_{occ}(d) + S_{unocc}(d)$. The modified version of segment cost function that concerns the occlusion situation is:

$$Cost(S, d) = \frac{1}{|S_{unocc}(d)|} \sum_{p \in S_{unocc}(d)} Cost(p, d) \tag{4}$$

5.3 Hierarchical Stereo Matching

Hierarchical stereo matching processes disparity level from foreground to background. Once a disparity level is recognized, the corresponding disparities in D_l and D_r are updated. Let d_{cur} denote the disparity of current processing level and $d_{min}(S)$ denote the disparity that gives the minimum matching cost using (4) for segment S . Consider current depth level’s candidate segment set $S_{candi}(d_{cur}) = \{S | d_{min}(S) = d_{cur}\}$. We claim that the element S in $S_{candi}(d_{cur})$ belongs to the d_{cur} level if and only if S satisfies consistency constraint. The consistency constraint of pixel is trivial, as defined in section 2. In the case of segment, the situation is a little more complex.

Definition 7 (Segment Consistency Constraint). Thinking of the ratio of a segment S between the count of unoccluded pixels satisfying pixel consistency

Algorithm 1: The matching algorithm for one disparity level

Input: The current disparity level
Output: The segments corresponding to the current disparity level

```

1 DealWithDisparityLevel( $d_{cur}$ )
2 {
3   Start from left image, find the set  $S_{candi}(d_{cur}) = \{S_l | d_{min}(S_l) = d_{cur}\}$ ;
4   If there is no appropriate segment  $S_l$  in  $S_{candi}(d_{cur})$  satisfying the consistency
   constraint, then:
5     Mark  $d_{cur}$  as invalid disparity; return;
6   For every  $S_l$  in  $S_{candi}(d_{cur})$ :
7     If  $S_l$  satisfies the consistency constraint, then:
8       DealWithSegment( $S_l$ );
9   }
10 DealWithSegment( $S$ )
11 {
12   Let  $I$  denote left or right image  $S$  belongs to and  $J$  denote the other image;
13   For every pixel  $p$  in  $S$ : Set  $D_I(p) = d_{cur}$ ;
14   Find the set  $S_{corr}(S) = \{S_J\}$  in the other image;
15   For every  $S_J$  in  $S_{corr}(S)$ :
16     If  $S_J$  satisfies the consistency constraint, then:
17       DealWithSegment( $S_J$ );
18   Else:
19     Split the pixels  $\{q\}$  corresponding to  $S$  from  $S_J$ ;
20     For every  $q$  in  $\{q\}$ : Set  $D_J(q) = d_{cur}$ ;
21 }

```

constraint and the total count of unoccluded pixels, we denote it as consistent ratio and define a threshold γ . We claim that the segment satisfies consistency constraint if and only if it's consistent ratio is greater than γ . Denote p as the pixel of left image, $q = p - d_{cur}$ as the pixel of right image, and $S(q)$ as the segment pixel q belonging to, we formulate segment consistency constraint as: $|\{p | p \in S_{unocc}(p) \wedge d_{min}(S(q)) = d_{cur}\}| / |S_{unocc}(p)| > \gamma$.

As described in the previous section, the foreground which has the maximum disparity is doubtless unoccluded. When initialization, the scene structure kept unknown, d_{cur} is set to be the maximum disparity, and segments which have the largest initial disparity and correspond with the consistency constraint are seemed as correct foreground. In the successive process d_{cur} is decreased by 1 each time until it is equal to the minimum disparity.

For every disparity level, if segment S_l in $S_{candi}(d_{cur})$ of left image corresponds with the consistency constraint we set d_{cur} as disparity of S_l . In general, the pixels corresponding to S_l in the right image belongs to several segments, which we denote as $S_{corr}(S_l) = \{S_r\}$. If S_r in $\{S_r\}$ satisfies the consistency constraint we set d_{cur} as it's disparity, otherwise, split the pixels corresponding to S_l from S_r and set their disparities d_{cur} . It is obviously that this process is left-right iterative. When candidate segments in both images satisfy the consistency constraint, the d_{cur} level process stops.

Especially if there is no appropriate segment S in $S_{candi}(d_{cur})$ satisfying consistency constraint the current disparity d_{cur} is invalid and should be eliminated. The matching algorithm for one disparity level is given in Algorithm. 1.

5.4 Undetermined Segments

There are still some undetermined segments after the step described in subsection 5.3 finished mainly due to ambiguous match. Consider the envelope composed of GCPs around the undetermined segment, it confines the possible disparities of the undetermined segment, which comprise of disparities of GCPs' on the envelope. When the disparity range is determined, a little more computation helps to resolve the remained problem.

For single undetermined pixel we offer a more quick method. Consider the GCPs on 8-neighbor direction. Among those eight disparities if some d 's count is greater than a threshold and the positions correspond to some distribution, we can set the undetermined pixel disparity d directly. Fig.4 demonstrates three valid distributions. There are 12 valid distributions by symmetry.

Especially, if the undetermined segment is surrounded by several segments on the same plane, we can compute each pixel's disparity on the undetermined segment directly using plane equation, as discussed in the next subsection.

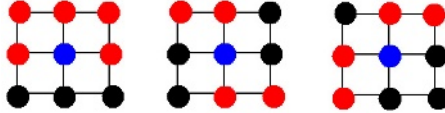


Fig. 4. Determining single undetermined pixel's disparity directly by 8-neighbor direction GCPs. GCPs with the same disparities are marked in red and undetermined center pixel in blue. Center pixel's disparity can be set as the red pixels' in the three cases.

5.5 Plane Fitting

For scenes in which there exist several planes such as "Venus" and "Sawtooth" in the Middlebury database[1], we tried fitting the planes to improve scene smoothness using plane equation (5) as [15,17] did:

$$d = ax + by + c \quad (5)$$

where x and y are pixel p 's image coordinates, a , b and c are plane parameters.

Disparities of pixels on the slant plane are changed gradually and integral disparities are no more fit here. In such cases we use semi-disparity cost function (6) to get the proper disparities. When the hierarchical algorithm processes d_{cur} is decreased by $1/2$ each time.

$$Cost(S, d + \frac{1}{2}) = \frac{1}{2}(Cost(S, d) + Cost(S, d + 1)) \quad (6)$$

Here we propose an efficient way to find segments on the same plane and to determine scene planes. When most segments are designated disparities, we random select a segment as seed segment and examine the neighboring segments' disparities. If the difference between two disparities of the seed segment and the neighboring segment is not greater than $1/2$, we mark the two segments a same plane ID. Then the marked neighboring segments are treated as seem segment individually and such operation repeats. If undetermined segment is surrounded by several segments marked as the same plane ID we set the undetermined segment that ID too. For each plane, the parameters a , b and c can be resolved by the least square solution from the linear system

$$A[a, b, c]^T = B \quad (7)$$

where each row of A is the $[x, y, 1]$ vector for all pixels in the plane, and each row of B is pixel's corresponding disparity d . Conversely, disparity of each pixel can be redesignated using (5) to improve smoothness.

6 Experimental Results

In this section we present experimental results on the Middlebury database [1]. They provide stereo images with ground truth, evaluation software, and comparison with other algorithms (<http://cat.middlebury.edu/stereo/>). This database has become a benchmark for dense stereo algorithm evaluation.

For all the experiments, we set $\lambda = 1$ and $\gamma = 0.8$. Fig.5 shows the result on four stereo pairs from the Middlebury database. Table 1 summarizes the results of evaluation. The algorithms are listed roughly in decreasing order of overall performance, and the minimum (best) value in each column is shown in red. We

Table 1. Middlebury stereo evaluation table. The first column lists algorithm name. The next 4 columns give percentage errors on the four scenes. Each of these four columns is broken into 2 or 3 subcolumns: the all, disc, and untex columns give the total error percentage everywhere, in the untextured areas, and near discontinuities, respectively.

Algorithm	Tsukuba			Sawtooth			Venus			Map	
	all	untex	disc	all	untex	disc	all	untex	disc	all	disc
Sym.BP+occl.	0.97	0.28	5.45	0.19	0.00	2.09	0.16	0.02	2.77	0.16	2.20
Patch-based	0.88	0.19	4.95	0.29	0.00	3.23	0.09	0.02	1.50	0.30	4.08
Segm.-based GC	1.23	0.29	6.94	0.30	0.00	3.24	0.08	0.01	1.39	1.49	15.46
Graph+segm.	1.39	0.28	7.17	0.25	0.00	2.56	0.11	0.02	2.04	2.35	20.87
Proposed	1.07	0.38	6.43	0.25	0.00	2.58	0.29	0.03	4.87	0.66	9.21
GC + mean shift	1.13	0.48	6.38	1.14	0.06	3.34	0.77	0.70	3.61	0.95	12.83
Segm.+glob.vis.	1.30	0.48	7.50	0.20	0.00	2.30	0.79	0.81	6.37	1.63	16.07
Belief prop.	1.15	0.42	6.31	0.98	0.30	4.83	1.00	0.76	9.13	0.84	5.27
Layered	1.58	1.06	8.82	0.34	0.00	3.35	1.52	2.96	2.62	0.37	5.24
2-pass DP	1.53	0.66	8.25	0.61	0.02	5.25	0.94	0.95	5.72	0.70	9.32

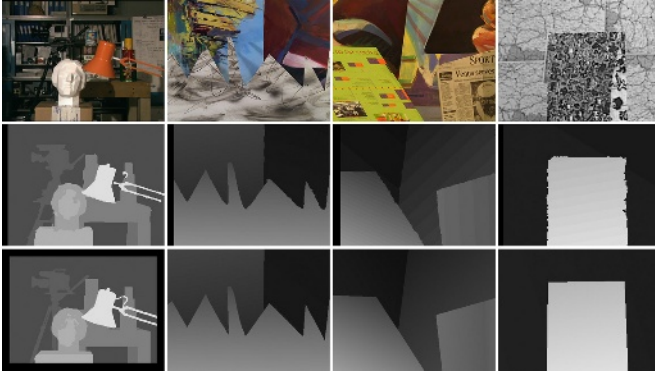


Fig. 5. Results on Middlebury datasets. From left to right order: Tsukuba, Sawtooth, Venus, Map. From top to down order: reference images, our result maps, the ground truth maps.

list only top 10 results(at the submission time, including all published methods). In fact the ranking gives just a rough idea of the performance of an algorithm and there is actually little difference between those algorithms. It is hard to come up with a "perfect" ranking function since no algorithm shows best performance on all four images. Our algorithm is in the bold and black face.

Besides the strong numerical and visual performance, another distinct advantage of the proposed algorithm is memory saving performance. When initialization, we allocate memory for all possible matches (p, d) and (S, d) to save corresponding costs. Whenever a disparity level is identified or an invalid disparity is eliminated, the memories allocated for this disparity are released. The total memory space our algorithm holds decrease dynamically with the process running until terminated.

7 Conclusions

In this paper, we present a new segment-based stereo method that handles occlusion and obtains disparity map from foreground to background. A robust, efficient and flexible hierarchical matching algorithm is developed. Experiments demonstrate that the performance of our approach is comparable to the state-of-the-art stereo algorithms on various datasets.

Acknowledgements

This work is supported by the National Basic Research Program of China(973 Program) under Grant No.2004CB719403 and the National High Technology Research and Development Program of China under Grant No.2004AA115120.

References

1. D.Scharstein, R.Szeliski: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision* **47** (2002) 7–42
2. M.Z.Brown, D., G.D.Hager: Advances in computational stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25** (2003) 993–1008
3. M.Okutomi, T.Kanade: A multiple baseline stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15** (1993) 353–363
4. A.Fusiello, V., E.Truccho: Efficient stereo with multiple windowing. In: *IEEE Computer Vision and Pattern Recognition*. (1997) 858–863
5. D.Geiger, B., A.Yuille: Occlusions and binocular stereo. In: *European Conference on Computer Vision*. (1992) 425–433
6. O.Veksler: Fast variable window for stereo correspondence using integral images. In: *IEEE Computer Vision and Pattern Recognition*. (2003) 556–564
7. S.Birchfield, C.Tomasi: Depth discontinuities by pixel-to-pixel stereo. *International Journal of Computer Vision* **35** (1999) 269–299
8. A.F.Bobick, S.S.Intille: Large occlusion stereo. *International Journal of Computer Vision* **33** (1999) 181–200
9. Minglun Gong, Yee Hong Yang: Fast stereo matching using reliability-based dynamic programming and consistency constraints. In: *IEEE International Conference on Computer Vision*. (2003) 610–617
10. T.H.Cormen, C.E.Leiserson, R., C.Stein: *Introduction to algorithms*. Higher Education Press and The MIT Press, Beijing (2002)
11. S.Roy: Stereo without epipolar lines: A maximum-flow formulation. *International Journal of Computer Vision* **34** (1999) 147–161
12. S.Birchfield, C.Tomasi: Multiway cut for stereo and motion with slanted surfaces. In: *IEEE International Conference on Computer Vision*. (1999) 489–495
13. Y.Boykov, O., R.Zabih: Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23** (2001) 1222–1239
14. Y.Boykov, V.Kolmogorov: An experimental comparison of min-cut / max-flow algorithms for energy minimization in vision. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **26** (2004) 1124–1137
15. H.Tao, H., R.Kumar: A global matching framework for stereo computation. In: *IEEE International Conference on Computer Vision*. (2001) 532–539
16. Yichen Wei, Long Quan: Region-based progressive stereo matching. In: *IEEE Computer Vision and Pattern Recognition*. (2004) 106–113
17. Li Hong, George Chen: Segment-based stereo matching using graph cuts. In: *IEEE Computer Vision and Pattern Recognition*. (2004) 74–81
18. D.Comaniciu, P.Meer: Mean shift: A robust approach toward feature space analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **24** (2002) 603–619

Gabor Feature Based Face Recognition Using Supervised Locality Preserving Projection

Zhonglong Zheng¹, Jianmin Zhao¹, and Jie Yang²

¹ Institute of Information Science and Engineering, Zhejiang Normal University, Jinhua, Zhejiang, China

zhonglong@zjnu.cn

² Institute of Image Processing and Pattern Recognition, Shanghai Jiaotong University, Shanghai, China

Abstract. This paper introduces a novel Gabor-based supervised locality preserving projection (GSLPP) method for face recognition. Locality preserving projection (LPP) is a recently proposed method for unsupervised linear dimensionality reduction. LPP seeks to preserve the local structure which is usually more significant than the global structure preserved by principal component analysis (PCA) and linear discriminant analysis (LDA). In this paper, we investigate its extension, called supervised locality preserving projection (SLPP), using class labels of data points to enhance its discriminant power in their mapping into a low dimensional space. The GSLPP method, which is robust to variations of illumination and facial expression, applies the SLPP to an augmented Gabor feature vector derived from the Gabor wavelet representation of face images. We performed comparative experiments of various face recognition schemes, including the proposed GSLPP method, principal component analysis (PCA) method, linear discriminant analysis (LDA) method, locality preserving projection method, the combination of Gabor and PCA method (GPCA) and the combination of Gabor and LDA method (GLDA). Experimental results on AR database and CMU PIE database show superior of the novel GSLPP method.

1 Introduction

In the past decades, there have been many methods proposed for dimensionality reduction ([1]-[7] and [15]). Two canonical forms of them are principal component analysis (PCA) and multidimensional scaling (MDS). Both of them are eigenvector methods aimed at modeling linear variability in the multidimensional space.

Recently, an unsupervised linear dimensionality reduction method, locality preserving projection (LPP), was proposed and applied to real datasets ([8]-[13]). LPP aims to preserve the local structure of the multidimensional structure instead of global structure preserved by PCA. In addition, LPP shares some similar properties compared with LLE such as a locality preserving character. However, their objective functions are totally different. LPP is the optimal linear

approximation to the eigenfunctions of the Laplace Beltrami operator on the manifold ([26]). LPP is linear and can deal with new data easily. In contrast, LLE is nonlinear and unclear how to evaluate test points.

In this paper, we describe a supervised variant of LPP, called the supervised locality preserving projection (SLPP) algorithm. Unlike LPP, SLPP projects high dimensional data to the embedded low space taking class membership relations into account. This allows obtaining well-separated clusters in the embedded space. It is worthwhile to highlight the discriminant power of SLPP by using class information besides inheriting the properties of LPP. Therefore, SLPP demonstrates powerful recognition performance when applied to some pattern recognition tasks. The GSLPP method for face recognition, which is robust to variations of illumination and facial expression, applies the SLPP to an augmented Gabor feature vector derived from the Gabor wavelet representation of face images. We performed comparative experiments of various face recognition schemes, including the proposed GSLPP method, principal component analysis (PCA) method, linear discriminant analysis (LDA) method, locality preserving projection method, the combination of Gabor and PCA method (GPCA) and the combination of Gabor and LDA method (GLDA).

2 Locality Preserving Projection

We can conclude that both PCA and LDA aim to preserve the global structure. In fact, the local structure is more important in many real cases. Locality preserving projection (LPP) is a linear approximation algorithm of the non-linear Laplacian Eigenmap for learning a locality preserving subspace ([26]). LPP aims to preserve the intrinsic geometry of the data and local structure. The objective function of LPP is defined as:

$$\min \sum_{ij} \|y_i - y_j\|^2 S_{ij} \tag{1}$$

where S is a symmetry similarity measure matrix. A possible way of defining such S is:

$$S_{ij} = \begin{cases} \exp(-\|x_i - x_j\|^2/t), & \|x_i - x_j\|^2 < \epsilon \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where $\epsilon > 0$ defines the radius of the local neighborhood. Here the selection of ϵ is somewhat like that of in LLE algorithm ([6]). The imposed constraint is $y^T D y = 1$ ([9]). Finally, the minimization problem reduces to the following form:

$$\begin{aligned} & \arg \min_W W^T X L X^T W \\ & \text{with } W^T X D X^T W = 1 \end{aligned} \tag{3}$$

where D is a diagonal matrix, $D_{ii} = \sum_j S_{ij}$. And $L = D - S$ is the Laplacian matrix ([3]). The transformation vector W_{LPP} is determined by the minimum eigenvalue solution to the generalized eigenvalue problem:

$$X L X^T W = \lambda X D X^T W \tag{4}$$

For more detailed information about LPP, please refer to [9][10][11][26].

3 Supervised Locality Preserving Projection

It is our motivation to combine locality preserving property with discriminant information to enhance the performance of LPP in pattern analysis ([14], [16]-[19], [26]). Being unsupervised, the original LPP does not make use of class membership relation of each point to be projected. To complement the original LPP, a supervised LPP has been proposed in the paper, called SLPP.

Let's rearrange the order of samples in the original dataset which will not affect the procedure of the algorithm. Suppose that the first M_1 columns of X are occupied by the data of the first class, the next M_2 columns are composed of the second class, etc., i.e. data of a certain class are compactly stored in X . This step is of benefit to simplify the explanation of SLPP algorithm. As a consequence, X is changed to be an orderly matrix which is composed of sub-matrices A_i of size $n \times M_i, i = 1, \dots, c$, where c is the number of classes. In the same manner, B_2, \dots, B_c are generated by repeating the same process. Then the nearest neighbors for each $x_j \in A_1$ are sought in A_1 only. When applied to all $x_j \in A_1$, the procedure leads to a construction of the matrix B_1 . When obtained B_1, \dots, B_c , the similarity measure matrix S is constructed by taking B_i as its diagonal structural elements, i.e.:

$$S = \begin{bmatrix} B_1 & & & \\ & B_2 & & \\ & & \dots & \\ & & & B_c \end{bmatrix} \tag{5}$$

To simplify the similarity computation between different points, we just set the weight equal to 1 if the two points belong to the same class. Therefore, B_i takes the following form:

$$B_i = \underbrace{\begin{bmatrix} 0 & 1 & \dots & 1 \\ 1 & 0 & & 1 \\ \vdots & & \ddots & \vdots \\ 1 & \dots & 1 & 0 \end{bmatrix}}_{M_i} \tag{6}$$

The algorithmic procedure of SLPP is then stated as:

- (1) Order Rearrangement. As we have described in paragraph 4 of this section, samples of a certain class are stored compactly in original samples matrix X after order rearrangement.
- (2) PCA Projection. For the sake of avoiding singularity of $XDXT^T$ and of reducing noise, we project X to its PCA subspace. X is still used to denote samples in PCA space, and the transformation matrix is denoted by W_{PCA} .
- (3) Computing Similarity Measure Matrix S . Because we have finished the order rearrangement of samples, B_i in Eq6 is easily computed. Then S is constructed by taking B_i as its diagonal structural elements.

(4) Eigenmap. Solve the generalized eigenvector problem:

$$X L X^T W = \lambda X D X^T W \tag{7}$$

Then the final transformation matrix from original sample space to the embedded feature space is:

$$W_{SLPP} = W_{PCAW} \tag{8}$$

where W is the solution of Eq7.

4 Gabor Wavelets

Marcelja and Dauman discovered that simple cells in the visual cortex can be modeled by Gabor functions ([28]). The 2D Gabor functions proposed by Daugman are local spatial band-pass filters that achieve the theoretical limit for conjoint resolution of information in the 2D spatial and 2D Fourier domains, that is, Gabor wavelets exhibit desirable characteristics of spatial locally and orientation selectivity. Donato et al. had recently shown through experiments that the Gabor wavelet representation gives better performance than other techniques for classifying facial actions ([27]).

The Gabor wavelets (kernels, filters) can be defined as:

$$\Psi_{\alpha,\beta}(z) = \frac{\|k_{\alpha,\beta}\|^2}{\sigma^2} e^{(-\|k_{\alpha,\beta}\|^2 \|z\|^2 / 2\sigma^2)} (e^{ik_{\alpha,\beta}z} - e^{-\sigma^2/2}) \tag{9}$$

where α and β define the orientation and scale of the Gabor kernels, $\|\cdot\|$ denotes the norm operator, $z = (x, y)$, and the wave vector $k_{\alpha,\beta}$ is defined as:

$$k_{\alpha,\beta} = k_{\beta} e^{i\phi_{\alpha}} \tag{10}$$

where $k_{\beta} = k_{max}/f^{\beta}$ and $\phi_{\alpha} = \pi\alpha/8$. k_{max} is the maximum frequency, and f is the spacing factor between kernels in the frequency domain.

Let $f(x, y)$ be the gray level distribution of the image. The Gabor wavelet representation of the image is the convolution of $f(x, y)$ with a series of Gabor kernels at different scales and orientation:

$$Y_{\alpha,\beta}(z) = f(z) * \Psi_{\alpha,\beta}(z) \tag{11}$$

where $\| * \|$ denotes the convolution operator, $Y_{\alpha,\beta}$ is the corresponding convolution result related to different orientation α and β . Applying the convolution theorem, it gives:

$$Y_{\alpha,\beta}(z) = \mathcal{F}^{-1}\{\mathcal{F}\{f(z)\}\mathcal{F}\{\Psi_{\alpha,\beta}(z)\}\} \tag{12}$$

where \mathcal{F} denoted the Fourier transform.

In order to encompass all frequency and locality information as much as possible, this paper, same as Liu [29], concatenated the all Gabor representations at

the five scales and eight orientations. Before the concatenation, $Y_{\alpha,\beta}(z)$ is down-sampled by a factor ρ to reduce the space dimension, and normalized to zero mean and unit variance. We then construct a vector out of the $Y_{\alpha,\beta}(z)$ by concatenating its rows (or columns). Now let $Y_{\alpha,\beta}^\rho(z)$ denote the normalized vector constructed from $Y_{\alpha,\beta}(z)$, the augmented Gabor feature vector Y^ρ is defined as:

$$Y^\rho = (Y_{\alpha,\beta}^\rho | \alpha = 0, \dots, 7; \beta = 0, \dots, 4) \quad (13)$$

5 Experimental Results of GSLPP

In this section, we applied the proposed GSLPP to the face recognition task, together with PCA, LDA, LPP, GPCA and GLDA. Before carrying out the experiments, each image will be transformed into its Gabor representation in Eq13 both in training process and in testing process. The face data sets include the well known AR data set and CMU PIE data set.

5.1 AR Face Dataset

AR face data set consists of 26 frontal images with different facial expressions, illumination conditions, and occlusions (sunglass and scarf) for 126 subjects (70 men and 56 women). Images were recorded in two different sessions separated by two weeks ([20]). Thirteen images were taken at the CVC under strictly controlled conditions in each session. The images taken from two sessions of one specific person are shown in Fig.1.



Fig. 1. Some samples of AR face data set

The size of the images in AR data set is 768×576 pixels, and each pixel is represented by 24 bits of RGB color values. We select 80 individuals (50 men and 30 women) for our experiment. 13 images in each session were all chosen for every subject. That is, there are 26 images per subject in our database. Totally, our database includes 2,080 images. Then these images were the input of a face detection system which combines skin-based method and boosting method ([30][31]). After face detection, the detected face images are converted to gray-scale and resized to 40×40 . In fact, there are still a very small number of faces not

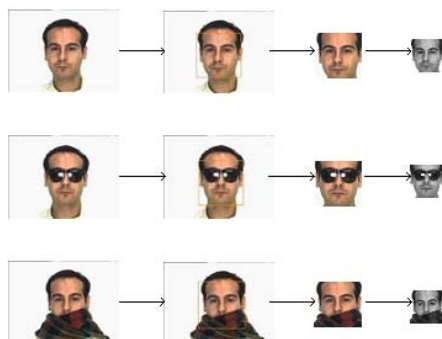


Fig. 2. Face detection and preprocessing of AR

detected by the detection system although it achieves over 98% detection rate on our database. For the purpose of experiment, these undetected face images were cropped manually. The flow chart of detection process is shown in Fig.2.

The face images wearing glasses and scarf in the first session and second session are contained in the training set and testing set, respectively. The left 14 images are divided randomly into two parts: 7 images for training and the other 7 for testing. That is, there are 1,040 images in the training set, and 13 images per subject. It is the same to the testing set. In terms of theoretical analysis, different method (PCA, LDA, and LPP) would result in different embedding feature space. In fact, the basis of the feature space is the eigenvectors of the corresponding method. Furthermore, we can display these eigenvectors as images. Because these images look like human faces, they are often called Eigenfaces, Fisherfaces, Laplacianfaces ([10][14][20]). For the eigenvectors of our SLPP, it may be called S-Laplacianfaces.

The S-Laplacianfaces derived from the training set are shown in Fig.3, together with Eigenfaces, Fisherfaces and Laplacianfaces. The nearest neighbor classifier is employed for classification. The recognition rates reach the best results with 121, 101, 79, 79, 109, 82 dimensions for PCA, GPCA, LDA, GLDA,

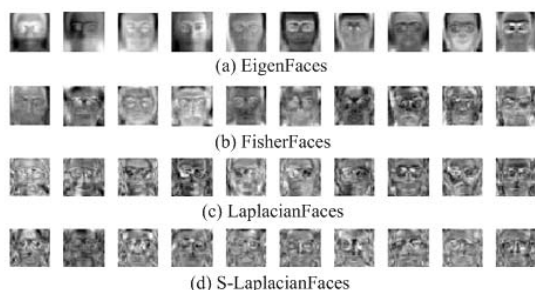
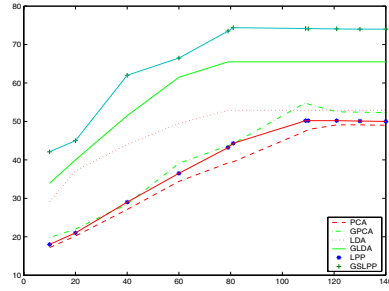


Fig. 3. Feature-faces of different algorithms on AR

Algorithm	Recognition Rate	Dimension
PCA	49.1%	121
GPCA	54.5%	101
LDA	52.9%	79
GLDA	65.5%	79
LPP	50.2%	109
GSLPP	74.3%	85

(a) the best recognition rates



(b) recognition rates vs. dimensions

Fig. 4. Experimental results on AR face data set

LPP and GSLPP, respectively. Fig.4 shows the recognition rates versus dimensionality reduction.

GSLPP method achieves the best recognition rate comparing with the other five methods, though the recognition accuracy obtained by all these methods is relatively low. One possible reason is that there are some occluded images in training set and testing set. These images severely deteriorate the determinant power of these methods besides the exaggerated expression of non-occluded images. But the most important we want to demonstrate is that GSLPP is more effective than the other five methods. And the experiment proves its good performances.

5.2 CMU PIE Face Dataset

The CMU PIE face dataset consists of 68 subjects with 41,368 face images as a whole. The face images were captured by 13 synchronized cameras and 21 flashes, under varying pose, illumination, and expression ([32]). In this subsection, we selected 40 subjects in the database and used 140 face images for each individual. 70 images are for training, and the other 70 images for testing. Still, faces are detected by the face detection system stated in section 5.1. The detected faces are converted to grayscale and resized to 40×40 , no other preprocessing. Some samples are shown in Fig.5. Totally, there are 2,800 images in the training set and the testing set, respectively.

For the sake of visualization, we illustrate S-Laplacianfaces derived from the training set, together with Eigenfaces, Fisherfaces and Laplacianfaces in Fig.6. Still, nearest neighbor classifier was adopted to perform the recognition task for its simplicity, though there are other classifiers for pattern recognition such as Neural Network [23], Bayesian [21], Support Vector Machine [22], etc.

The recognition rates approach the best with 115, 95, 39, 39, 107 and 77 dimensions for PCA, GPCA, LDA, GLDA, LPP and GSLPP, respectively. Fig.7 illustrates the recognition rates versus dimensionality reduction.



Fig. 5. Some samples of CMU-PIE face data set

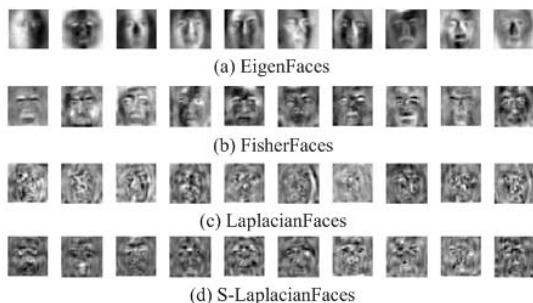
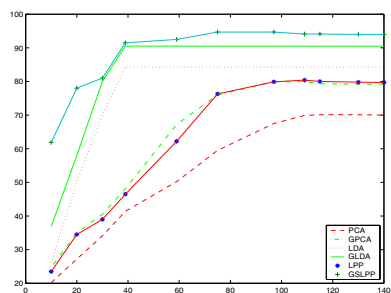


Fig. 6. Feature-faces of different algorithms on CMU-PIE

Algorithm	Recognition Rate	Dimension
PCA	70.1%	115
GPCA	79.9%	95
LDA	84.3%	39
GLDA	90.5%	39
LPP	80.2%	107
GSLPP	94.7%	77

(a) the best recognition rates



(b) the recognition rates vs. dimensions

Fig. 7. Experimental results on CMU-PIE face data set

6 Discussion and Future Work

It is worthwhile mentioning that GSLPP takes advantage of more training samples, which is important to learn an efficient and effective embedding space representing the non-linear structure of original patterns. Sometimes there might be only one training sample available for each class. In such a case, GSLPP can not work since the similarity matrix is a nought matrix. How to overcome such problem is one of our future works.

In fact, it is not always the case that we can obtain all the class information in pattern analysis. Sometimes we only have unlabeled samples or partially labeled samples. Therefore, another extension of our work is to consider how to use these unlabeled samples for discovering the manifold structure and, hence, improving

the classification performance ([10][25]). We are now exploring these problems with full enthusiasm.

References

1. I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, New York, 1986.
2. H.Klock and J. Buhmann. Data visualization by multidimensional scaling: a deterministic annealing approach. *Pattern Recognition*, 33(4):651-669, 1999
3. M. Belkin and P. Niyogi. Laplacian Eigenmaps and Spectral Techniques for Embedding and Clustering. *Proc. Conf. Advances in Neural Information Processing System 15*, 2001
4. D.D.Lee, and H.S.Seung, "Learning the parts of objects by non-negative matrix factorization", *Nature* 401, 788-791 (1999)
5. M.S. Barlett, H.M. Ladesand, and T.J. Sejnowsky, "Independent component representations for face recognition", *Proc. SPIE* 3299, 528-539 (1998)
6. S. T. Roweis, L. K. Saul. "Nonlinear dimensionality reduction by locally linear embedding", *Science*, 290(5500), pp.2323-2326, 2000
7. J. B. Tenenbaum et al., "A global geometric framework for nonlinear dimensionality reduction", *Science* 290(5500), pp. 2319-2323, 2000
8. Zhonglong Zheng, Jie Yang. Extended LLE with Gabor Wavelet for Face Recognition. *The 17th Australian Joint Conference on Artificial Intelligence 6-10th December 2004, Cairns Australia*
9. X. He and P. Niyogi. Locality Preserving Projections. *Proc. Conf. Advances in Nerual Information Processing Systems*, 2003
10. X. He, Shicheng Yan, et al. Face Recognition Using Laplacianfaces. *IEEE trans. On PAMI*, vol.27, No.3, 328:340, 2005
11. Xin Zheng, Deng Cai, Xiaofei He, Wei-Ying Ma, and Xueyin Lin. Locality Preserving Clustering for Image Database. *ACM conference on Multimedia 2004, Oct 10-16, 2004, New York City*
12. Xiaofei He, Shuicheng Yan, Yuxiao Hu, and Hong-Jiang Zhang. Learning a Locality Preserving Subspace for Visual Recognition. *IEEE International Conference on Computer Vision (ICCV 2003)*, Nice, France, 2003
13. D. de Ridder and R. P. W. Duin. Locally linear embedding for classification. *Technical Report PH-2002-01*, Pattern Recognition Group, Dept.of Imaging Science & Technology, Delft University of Technology, Delft, Netherlands, 2002
14. P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman. Eigenfaces vs. Fisherfaces: Recognition Using Class Specific Linear Projection. *IEEE tran. On PAMI*, vol.19, No.7, 711:721, 1997.
15. W, Zhao, R. Chellappa, and N. Nandhakumar. Empirical Performance Analysis of Linear Discriminant Classifiers. *Proc. Computer Vision and Pattern Recognition*, 164:169, 1998
16. D. L. Swets and J. Weng. Using Discriminant Eigenfeatures for Image Retrieval. *IEEE trans. On PAMI*, vol.18, No.8, 831:836, 1996.
17. H. Cevikalp, M. Neamtu, et al. Discriminant Common Vectors for Face Recognition. *IEEE trans. On PAMI*, vol.27, No.1, 4:13, 2005
18. T. K. Kim, and J. Kittler. Locally Linear Discriminant Analysis for Multimodally Distributed Classes for Face Recognition with a Single Model Image. *IEEE trans. On PAMI*, vol.27, No.3, 318:327, 2005

19. Jian Yang, Alejandro F. Frangi, Jing-yu Yang, et al. KPCA Plus LDA: A complete Kernel Fisher Discriminant Framework for Feature Extraction and Recognition. *IEEE trans. On PAMI*, vol.27, No.2, 2005
20. M. Turk, A. Pentland, "Eigenfaces for recognition", *Journal of Cognitive neuroscience*, vol. 3, pp. 71-86, 1991
21. Ira Cohen, Nicu Sebe, Fabio G. Cozman, Marcelo C. Cirelo, Thomas S. Huang. Learning Bayesian Network Classifiers for Facial Expression Recognition with both Labeled and Unlabeled data. *IEEE conference on Computer Vision and Pattern Recognition 2003*
22. Arnulf B. A. Graf, Alexander J. Smola, and Silvio Borer. Classification in a normalized feature space using support vector machines. *IEEE trans. On PAMI*, vol.14, No.3, 597:605, 2003
23. Rui-Ping Li, Masao Mukaidono, and I. Burhan Turksen. A fuzzy neural network for pattern classification and feature selection. *Fuzzy Sets and systems*, 130 (2002) 101:108
24. A.M. Martinez and R. Benavente, "The AR face database," *CVC Tech. Report #24*, 1998
25. I. Cohen, F. G. Cozman, N. Sebe, M. C. Cirelo, and T. S. Huang. Semi-supervised Learning of Classifiers: Theory, Algorithms, and Their application to Human-Computer Interaction. *IEEE trans. On PAMI*, vol.26, No.12, 1553:1567, 2004
26. Wanli Min, Ke Lu, and Xiaofei He. Locality preserving projection. *Pattern Recognition Journal*, Volume 37, Issue 4, Pages 781-788, 2004
27. G. Donato, M. S.Bartlett, J. C. Hager, & al., "Classifying facial actions", *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 21, pp. 974-989, 1999
28. J. G. Daugman. Two-dimensional spectral analysis of cortical receptive field profiles. *Vis. Res.*, vol.20, pp.847-856, 1980
29. C.Liu, H. Wechsler, "A Gabor feature classifier for face recognition", *Proc. 8th IEEE Int. Conf. Computer Vision*, Vancouver, BC, Canada, July 9-12, 2001
30. Rein-Lien Hsu, Mohamed and Anil K.Jain, "Face Detection in Color Images", *IEEE Trans. on PAMI* 24, 696-706 (2002)
31. Viola, P., and Jones, M., "Rapid object detection using a boosted cascade of simple features", In: *Proc. Conf. Computer Vision and Pattern Recognition*. Kauai, HI, USA 1, 511-518 (2001)
32. T. Sim, S. Baker, and M. Bsat, "The CMU Pose, Illumination, and Expression (PIE) Database", *Proc. IEEE Int'l Conf. Automatic Face and Gesture Recognition*, (2002)

Alternative Fuzzy Clustering Algorithms with L_1 -Norm and Covariance Matrix

Miin-Shen Yang¹, Wen-Liang Hung², and Tsiung-Iou Chung¹

¹ Department of Applied Mathematics, Chung Yuan Christian University,
Chung-Li 32023, Taiwan
msyang@math.cycu.edu.tw

² Department of Applied Mathematics, National Hsinchu University of Education,
Hsin-Chu, Taiwan
wlhung@mail.nhcue.edu.tw

Abstract. In fuzzy clustering, the fuzzy c -means (FCM) algorithm is the best known and most used method. Although FCM is a very useful method, it is sensitive to noise and outliers so that Wu and Yang (2002) proposed an alternative FCM (AFCM) algorithm. In this paper, we consider the AFCM algorithms with L_1 -norm and fuzzy covariance. These generalized AFCM algorithms can detect elliptical shapes of clusters and also robust to noise and outliers. Some numerical experiments are performed to assess the performance of the proposed algorithms. Numerical results clearly indicate the proposed algorithms to be superior to the existing methods.

1 Introduction

Cluster analysis is a tool for clustering a data set into groups of similar characteristic. The conventional (hard) clustering methods restrict each point of the data set to exactly one cluster. Since Zadeh [12] proposed fuzzy sets, which produced the idea of partial membership described by a membership function, fuzzy clustering has been successfully applied in various areas (see Bezdek [1], Yang [10] and Hoppner et al. [5]). In the literature on fuzzy clustering, the fuzzy c -means (FCM) algorithm is the most well-known and used method.

Although FCM is a very useful clustering method, it has some drawbacks. For example it is sensitive to noise and outliers. To overcome the drawbacks of FCM, Wu and Yang [9] proposed an alternative FCM (AFCM) clustering algorithm. Because there are many varieties of generalized FCM algorithms, Yu and Yang [11] had recently unified these FCM varieties into a generalization model.

In this paper, we only focus on generalizing AFCM with L_1 norm and also with a covariance matrix. We know that the FCM clustering algorithm had been extended with L_1 norm by Jajuga [6], Bobrowski and Bezdek [2] and Hathaway et al. [4] and also extended with a covariance matrix by Gustafson and Kessel [3] and Krishnapuram and Kim [8].

The remainder of this paper is organized as follows. In Section 2, the FCM clustering algorithm with L_1 norm is reviewed. We then proposed the AFCM

algorithm with L-1 norm. In Section 3, the FCM clustering algorithm with a covariance matrix is reviewed. We then proposed the AFCM algorithm with a covariance matrix. Numerical examples are given and comparisons are made between the proposed algorithms and the existing methods in Section 4. Conclusions and remarks will be stated in Section 5.

2 AFCM with L1-Norm

Let $X = \{x_1, \dots, x_n\}$ be a data set in an s -dimensional Euclidean space R^s with its norm denoted by $\|\cdot\|$. For a given c , $2 \leq c < n$, $a = \{a_1, a_2, \dots, a_c\}$ denotes the cluster centers where $a_i \in R^s$. Let $\mu = \{\mu_{ij}\}_{c \times n} \in M_{fcn}$ be fuzzy c -partitions, where

$$M_{fcn} = \left\{ \mu = [\mu_{ij}]_{cn} \mid \forall i, \forall j, \mu_{ij} \geq 0, \sum_{i=1}^c \mu_{ij} = 1, n > \sum_{j=1}^n \mu_{ij} > 0 \right\} \tag{1}$$

Then the FCM objective function is defined as follows:

$$J_{FCM}(\mu, a) = \sum_{j=1}^n \sum_{i=1}^c (\mu_{ij})^m \|x_j - a_i\|^2 \tag{2}$$

where the weighting exponent $m > 1$ presents the degree of fuzziness. Thus, the FCM algorithm is iterated through the necessary conditions for minimizing $J_{FCM}(\mu, a)$ with the following update equations:

$$a_i = \frac{\sum_{j=1}^n \mu_{ij}^m x_j}{\sum_{j=1}^n \mu_{ij}^m}, \quad i=1,2,\dots,c \tag{3}$$

$$\mu_{ij} = \frac{\|x_j - a_i\|^{-2}}{\sum_{k=1}^c \|x_j - a_k\|^{-2}}, \quad i=1,2,\dots,c; j=1,2,\dots,n. \tag{4}$$

Jajuga [6] first replaced the L_2 norm $\|x_j - a_i\|$ with L_1 norm by creating the following FCM-L1 objective function:

$$J_{FCM-L_1}(\mu, a) = \sum_{j=1}^n \sum_{i=1}^c \mu_{ij}^m d_{ij} \quad \text{with} \quad d_{ij} = \sum_{k=1}^s |x_{jk} - a_{ik}|. \tag{5}$$

Bobrowski and Bezdek [2] and Hathaway et al. [4] then extended to L_p and L_∞ norms. Kersten [7] proposed the so-called fuzzy c -medians that had a similar

algorithm as Bobrowski and Bezdek [2]. To present the algorithm, we only consider $s = 1$. That is, $J_{FCM-L_1}(\mu, a) = \sum_{j=1}^n \sum_{i=1}^c \mu_{ij}^m |x_j - a_i|$. By Lagrangian method, we can get the update equation for the fuzzy c-partitions as follows:

$$\mu_{ij} = |x_j - a_i|^{-\frac{1}{m-1}} / \sum_{k=1}^c |x_j - a_k|^{-\frac{1}{m-1}}, \quad i=1,2,\dots,c; j=1,2,\dots,n. \tag{6}$$

However, the update equation for the cluster center a_i needs to take the derivative of the absolute value $|x_j - a_i|$. In this case, we may order the data set as $x_{q(1)} \leq x_{q(2)} \leq \dots \leq x_{q(n)}$ with an ordering function $q(l), l=1, \dots, n$ so that

the FCM-L1 objective function becomes $J_{FCM-L_1}(\mu, a) = \sum_{l=1}^n \sum_{i=1}^c \mu_{iq(l)}^m |x_{q(l)} - a_i|$.

Thus, the derivative of $J_{FCM-L_1}(\mu, a)$ with respect to a_i will be as follows:

$$dJ_{FCM-L_1} = -\sum_{l=1}^n (\mu_{iq(l)})^m \text{sign}^+(x_{q(l)} - a_i) \quad \text{with} \quad \text{sign}^+(z) = \begin{cases} 1, & z \geq 0 \\ -1, & z < 0 \end{cases}.$$

To solve the equation $dJ_{FCM-L_1} = 0$, i.e., $\sum_{l=1}^n (\mu_{iq(l)})^m \text{sign}^+(x_{q(l)} - a_i) = 0$,

we need to find an optimal r such that $\sum_{l=1}^r (\mu_{iq(l)})^m - \sum_{l=r+1}^n (\mu_{iq(l)})^m = 0$. We can get

the update form for the cluster center a_i as follows: Let $S = -\sum_{l=1}^n \mu_{il}^m$ and $r = 0$. If $S < 0$, we set $r = r + 1$ and let $S = S + 2\mu_{iq(r)}^m$. If $S \geq 0$, then stop.

According to this method, we give the FCM-L1 clustering algorithm as follows:

FCM-L1 Algorithm

Step 1: Fix $2 \leq c \leq n$ and $\epsilon > 0$, and let $k = 1$.

Give initials $a^{(0)} = \{a_1^{(0)}, \dots, a_c^{(0)}\}$.

Step 2: Compute the fuzzy c-partitions $\mu^{(k)}$ with $a^{(k-1)}$ using Eq. (6).

Step 3: Order the data set with $x_{q(1)} \leq x_{q(2)} \leq \dots \leq x_{q(n)}$.

Let $S = -\sum_{l=1}^n \mu_{il}^m$ and $r = 0$.

Step 4: Using $S = S + 2\mu_{iq(l)}^m$ and $r = r + 1$.

IF $S < 0$, update $a^{(k)} = x_{q(r)}$.

Step 4: Compare $a^{(k)}$ to $a^{(k-1)}$ in a convenient norm $\|\cdot\|$.

IF the norm of $a^{(k)}$ and $a^{(k-1)}$ is less than ε , STOP

ELSE $k = k + 1$ and return to step 2.

To consider a robust metric measure based on the robust statistic and the influence function, Wu and Yang [9] proposed an exponential distance $d(x_j, a_i) = (1 - \exp(-\beta \|x_j - a_i\|^2))^{1/2}$ to replace the Euclidean distance $\|x_j - a_i\|$ in the FCM objective function and called it an alternative FCM (AFCM) clustering algorithm. The AFCM objective function was defined as

$J_{AFCM}(\mu, a) = \sum_{j=1}^n \sum_{i=1}^c (\mu_{ij})^m (1 - \exp(-\beta \|x_j - a_i\|^2))$. The necessary conditions for minimizing $J_{AFCM}(\mu, a)$ are the following update equations:

$$\mu_{ij} = \left[1 - \exp(-\beta \|x_j - a_i\|^2) \right]^{-\frac{1}{m-1}} / \sum_{k=1}^c \left[1 - \exp(-\beta \|x_j - a_k\|^2) \right]^{-\frac{1}{m-1}} \quad (7)$$

$$a_i = \frac{\sum_{j=1}^n \mu_{ij}^m \exp(-\beta \|x_j - a_i\|^2) x_j}{\sum_{j=1}^n \mu_{ij}^m \exp(-\beta \|x_j - a_i\|^2)}, \quad i = 1, \dots, c, \quad j = 1, \dots, n. \quad (8)$$

We now consider the AFCM-L1 algorithm by replacing the exponential distance $d(x_j, a_i) = (1 - \exp(-\beta \|x_j - a_i\|^2))^{1/2}$ with $d(x_j, a_i) = (1 - \exp(-\beta |x_j - a_i|))$. Thus, we have the following AFCM-L1 objective function:

$$J_{AFCM-L1}(\mu, a) = \sum_{j=1}^n \sum_{i=1}^c (\mu_{ij})^m (1 - \exp(-\beta |x_j - a_i|)). \quad (9)$$

The update equation for the fuzzy c -partitions as follows:

$$\mu_{ij} = \left[1 / (1 - \exp(-\beta |x_j - a_i|)) \right]^{1/(m-1)} / \sum_{k=1}^c \left[1 / (1 - \exp(-\beta |x_j - a_k|)) \right]^{1/(m-1)} \quad (10)$$

Similarly, we order the data set as $x_{q(1)} \leq x_{q(2)} \leq \dots \leq x_{q(n)}$ for an ordering function $q(l), l = 1, \dots, n$ so that

$$J_{AFCM-L1} = \sum_{i=1}^c \sum_{l=1}^n (\mu_{iq(l)})^m \left\{ 1 - \exp(-\beta |x_{q(l)} - a_i|) \right\}. \text{ Thus,}$$

$$dJ_{AFCM-L1}(\mu, a) = - \sum_{i=1}^c \sum_{l=1}^n (\mu_{iq(l)})^m \beta \exp(-\beta |x_{q(l)} - z_i|) \text{sign}^+(x_{q(l)} - z_i) w \text{ with}$$

$$\text{sign}^+(z) = \begin{cases} 1, & z \geq 0 \\ -1, & z < 0 \end{cases}. \text{ The AFCM-L1 algorithm is given as follows:}$$

AFCM-L1 Algorithm

Step 1: Fix $2 \leq c \leq n$ and $\epsilon > 0$, and let $k = 1$.

Give initials $a^{(0)} = \{a_1^{(0)}, \dots, a_c^{(0)}\}$.

Step 2: Compute the fuzzy c-partitions $\mu^{(k)}$ with $a^{(k-1)}$ using Eq. (10).

Step 3: Order the data set with $x_{q(1)} \leq x_{q(2)} \leq \dots \leq x_{q(n)}$.

$$\text{Let } S = \sum_{i=1}^c \sum_{l=1}^n (\mu_{iq(l)})^m \beta \exp(-\beta |x_{q(l)} - a_i|) \text{ and } r = 0.$$

Step 4: Using $S = S + 2(\mu_{iq(l)})^m \beta \exp(-\beta |x_{q(l)} - a_i|)$ and $r = r + 1$.

IF $S < 0$, update $a^{(k)} = x_{q(r)}$.

Step 4: Compare $a^{(k)}$ to $a^{(k-1)}$ in a convenient norm $\|\cdot\|$.

IF the norm of $a^{(k)}$ and $a^{(k-1)}$ is less than ϵ , STOP

ELSE $k = k + 1$ and return to step 2.

3 AFCM with a Covariance Matrix

Gustafson and Kessel [3] considered the effect of different cluster shapes except for spherical shape by replacing the Euclidean distance $d(x_j, a_i) = \|x_j - a_i\|$ in FCM with the Mahalanobis distance $d(x_j, a_i) = \|x_j - a_i\|_{A_i}^2 = (x_j - a_i)^T A_i (x_j - a_i)$ where A_i is a positive definite $s \times s$ covariance matrix and its determinate $\det(A_i) = \rho_i$ is a fixed constant. We call this extension as FCM-cov. Thus, the FCM-cov objective function is given as

$$J_{FCM-cov}(u, a, A) = \sum_{j=1}^n \sum_{i=1}^c \mu_{ij}^m \|x_j - a_i\|_{A_i}^2 \tag{11}$$

where $\mu \in M_{fcn}$, $a = (a_1, \dots, a_c) \in R^{cs}$ and $A = \{A_1, \dots, A_c\}$ for which A_i is positive definite with $\det(A_i) = \rho_i$. The necessary conditions for minimizing $J_{FCM-cov}(u, a, A)$ are the following update equations:

$$a_i = \sum_{j=1}^n \mu_{ij}^m x_j / \sum_{j=1}^n \mu_{ij}^m, \tag{12}$$

$$\mu_{ij} = \left(\sum_{k=1}^c \left\| x_j - a_i \right\|_{A_i}^{2/(m-1)} / \left\| x_j - a_k \right\|_{A_k}^{2/(m-1)} \right)^{-1}, \quad \begin{matrix} i = 1, \dots, c, \\ j = 1, \dots, n, \end{matrix} \tag{13}$$

with $A_i = (\rho_i \det(S_i))^{1/s} S_i^{-1}$, $S_i = \sum_{j=1}^n \mu_{ij}^m (x_j - a_i)(x_j - a_i)^T, i = 1, \dots, c$. Therefore, the FCM-cov algorithm is summarized as follows:

FCM-cov Algorithm

Step 1: Fix $2 \leq c \leq n$ and $\epsilon > 0$, and let $s = 1$.

Give initials $\mu^{(0)} = (\mu_1^{(0)}, \dots, \mu_c^{(0)})$.

Step 2: Compute the cluster centers $a^{(s)}$ with $\mu^{(s-1)}$ using Eq. (12).

Step 3: Update $\mu^{(s)}$ with $a^{(s)}$ using Eq. (13).

Step 4: Compare $\mu^{(s)}$ to $\mu^{(s-1)}$ in a convenient matrix norm $\| \cdot \|$.

IF $\| \mu^{(s)} - \mu^{(s-1)} \| < \epsilon$, STOP

ELSE $s = s + 1$ and return to step 2.

We mention that the FCM-cov algorithm became an important extended type of FCM. Krishnapuram and Kim [8] discussed more about the FCM-cov algorithm with a new variation.

Next, we consider the AFCM-cov by replacing the distance $d(x_j, a_i) = \|x_j - a_i\|_{A_i}^2 = (x_j - a_i)^T A_i (x_j - a_i)$ with the exponential distance $d(x, y) = 1 - \exp(-\beta(x_i - y_j)^T A_j (x_i - y_j))$. Thus, the AFCM-cov objective function is $J_{AFCM-cov} = \sum_{i=1}^c \sum_{j=1}^n (\mu_{ij})^m \{1 - \exp(-\beta(x_j - z_i)^T A_i (x_j - z_i))\}$.

Similarly, the necessary conditions for minimizing $J_{AFCM-cov}(u, a, A)$ are the following update equations:

$$\mu_{ij} = \frac{\left[1 / \left(1 - \exp\left(-\beta(x_j - a_i)^T A_i(x_j - a_i)\right)\right)\right]^{1/(m-1)}}{\sum_{k=1}^c \left[1 / \left(1 - \exp\left(-\beta(x_j - a_i)^T A_i(x_j - a_i)\right)\right)\right]^{1/(m-1)}} \tag{14}$$

$$a_i = \frac{\sum_{j=1}^n (\mu_{ij})^m x_j \exp\left(-\beta(x_j - a_i)^T A_i(x_j - a_i)\right)}{\sum_{j=1}^n (\mu_{ij})^m \exp\left(-\beta(x_j - a_i)^T A_i(x_j - a_i)\right)} \tag{15}$$

where $A_i^{-1} = \left(|A_i^{-1}| / |p|\right)^{\frac{1}{n}} p$ and

$$p = \sum_{j=1}^n (\mu_{ij})^m \beta(x_j - a_i)(x_j - a_i)^T \exp\left[-\beta(x_j - a_i)^T A_i(x_j - a_i)\right]$$

Note that, since a_i in Eq. (15) cannot be directly solved we need to use an iterative method to achieve it. Therefore, the AFCM-cov algorithm can be described as follows:

AFCM-cov Algorithm

Step 1: Let $h(a, \mu)$ be the right term of Eq. (15).

Fix $2 \leq c \leq n$ and $\epsilon > 0$, and let $s = 1$.

Give initials $a^{(0)} = \{a_1^{(0)}, \dots, a_c^{(0)}\}$.

Step 2: Compute the fuzzy c-partitions $\mu^{(s)}$ with $a^{(s-1)}$ using Eq. (14).

Step 3: Update $a^{(s)}$ with $a^{(s)} = h(a^{(s-1)}, \mu^{(s)})$ using Eq. (15).

Step 4: Compare $a^{(s)}$ to $a^{(s-1)}$ in a convenient matrix norm $\|\cdot\|$.

IF $\|a^{(s)} - a^{(s-1)}\| < \epsilon$, STOP

ELSE $s=s+1$ and return to step 2.

4 Numerical Examples and Comparisons

To assess the performance of AFCM-L1, FCM, AFCM and FCM-L1, the mean-squared error (MSE) and CPU time are calculated from the Monte Carlo experiments from 1000 replications of the sample for each case. We consider the mixture normal and Cauchy models in order to examine the effect of tail probability on the MSE and CPU time. Note that all implemented algorithms have the same initials in each

experiment. The fuzziness index m for all algorithms are chosen with $m = 2$. We mention that Yu and Yang [11] have recently investigated the theoretical selection of m for different fuzzy clustering algorithms such as FCM and AFCM.

Table 1. MSE and CPU time of FCM, AFCM, FCM-L1 and AFCM-L1

Algorithms		(a) Normal mixture	(b) Cauchy mixture
AFCM-L1	MSE	0.020	0.006
	CPU	2.978 second	3.744 second
AFCM	MSE	0.050	4.358
	CPU	0.739 second	0.910 second
FCM-L1	MSE	0.089	0.031
	CPU	2.118 second	3.080 second
FCM	MSE	0.088	96.624
	CPU	0.615 second	1.680 second

Example 1. We generate a sample of size 500 from (a) the normal mixture $0.5N(0,1) + 0.5N(5,1)$, and (b) the Cauchy mixture $0.5Cauchy(0,0.5) + 0.5Cauchy(5,0.5)$. The results are shown in Table 1. In the normal mixture case, AFCM-L1 has good accuracy but it needs more CPU time. AFCM has better MSE and CPU time than FCM-L1. On the other hand, the accuracy of AFCM-L1 performs very well in the Cauchy mixture case. However, FCM-L1 has better MSE than AFCM. It illustrates that the accuracy of AFCM is affected by the tail probability.

Table 2. MSE of FCM, AFCM, FCM-L1 and AFCM-L1

r (scale)	0.3	0.5	0.7	0.9	1
AFCM-L1	0.002	0.006	0.013	0.028	0.148
AFCM	0.034	4.194	19.39	69.04	125.2
FCM-L1	0.010	0.031	0.062	0.111	0.212
FCM	28.28	96.12	120.1	164.8	311.2

Example 2. To see the effect of heavy tail on FCM, AFCM, FCM-L1 and AFCM-L1, we generate a sample of size 500 from a mixture of Cauchy distribution $0.5Cauchy(0, r) + 0.5Cauchy(5, r)$ with a scale parameter r . The results are shown in Table 2. From Table 2, we see that MSEs are increasing as r being increasing. The reason is that data clusters are more separate when r is small. Furthermore, AFCM-L1 produces the smallest MSE. It reflects that AFCM-L1 is able to tolerate the heavy tail distributions. But the performance of FCM is heavily affected by the

heavy tail distributions. It is interesting that the performance of AFCM and FCM-L1 is similar as $r=0.3$. But r becomes large, there is an inordinate difference in MSE. It also means that the heavy tail distributions have heavily disturbed the accuracy of AFCM.

Next, we consider the effect of different cluster shapes on FCM, AFCM, FCM-cov and AFCM-cov.

Example 3. We consider a sample of size 100 from a 2-variate normal mixture $0.5N(\mu_1, \Sigma_1) + 0.5N(\mu_2, \Sigma_2)$ with $\mu_1=(0,0)$, $\mu_2=(5,0)$, $\Sigma_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 1 \end{bmatrix}$ and

$\Sigma_2 = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$. The clustering results are shown in Fig. 1. In this figure, the “o” represents cluster 1 and the “+” represents cluster 2. From Fig. 1, we see that FCM-cov and AFCM-cov classified these two clusters without any misclassified data. But FCM and AFCM have an inaccurate clustering result with 2 misclassified data.

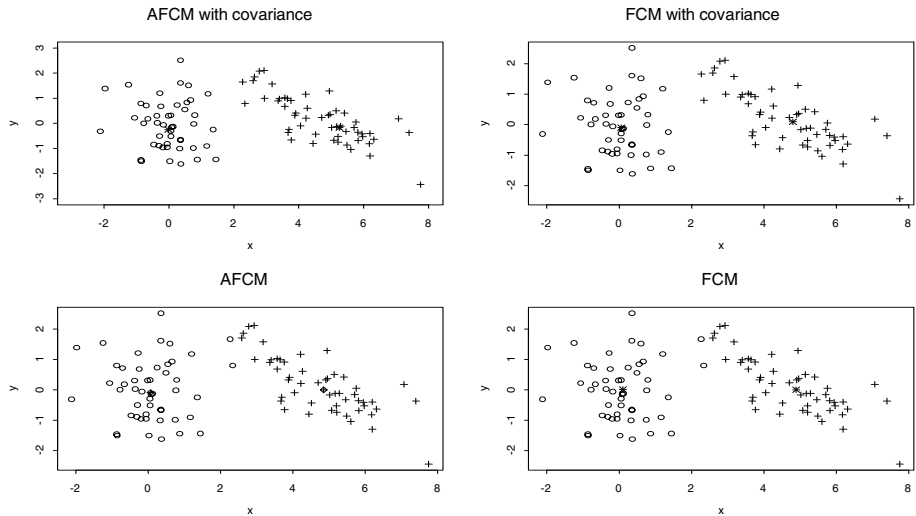


Fig. 1. Clustering results of FCM, AFCM, FCM-cov and AFCM-cov

Example 4. In this example, we generate a sample of size 100 from a 2-variate normal mixture $0.5N(\mu_1, \Sigma_1) + 0.5N(\mu_2, \Sigma_2)$ with $\mu_1=(0,0)$, $\mu_2=(4,0)$,

$\Sigma_1 = \begin{bmatrix} 0.3 & 0 \\ 0 & 0.3 \end{bmatrix}$ and $\Sigma_2 = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}$. The clustering results are shown

in Fig. 2. From Fig. 2, we see that AFCM-cov classified these two clusters without any misclassified data. But FCM, AFCM, FCM-L1 have an inaccurate

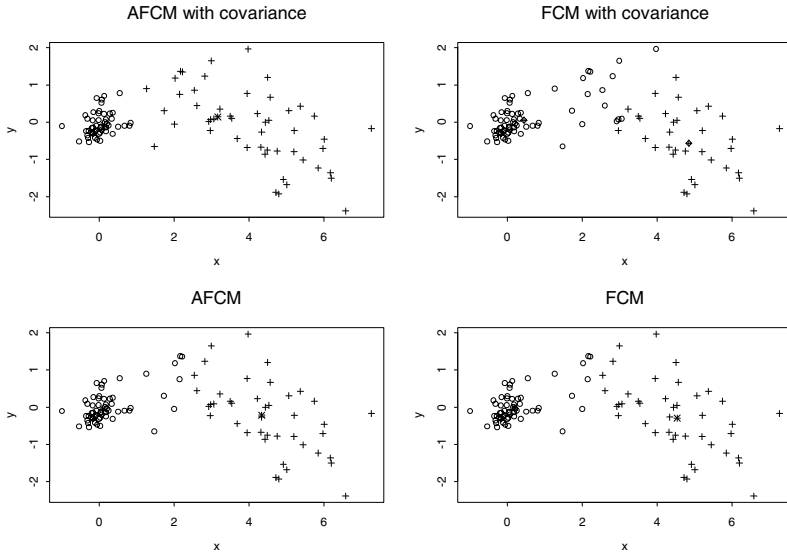


Fig. 2. Clustering results of FCM, AFCM, FCM-cov and AFCM-cov

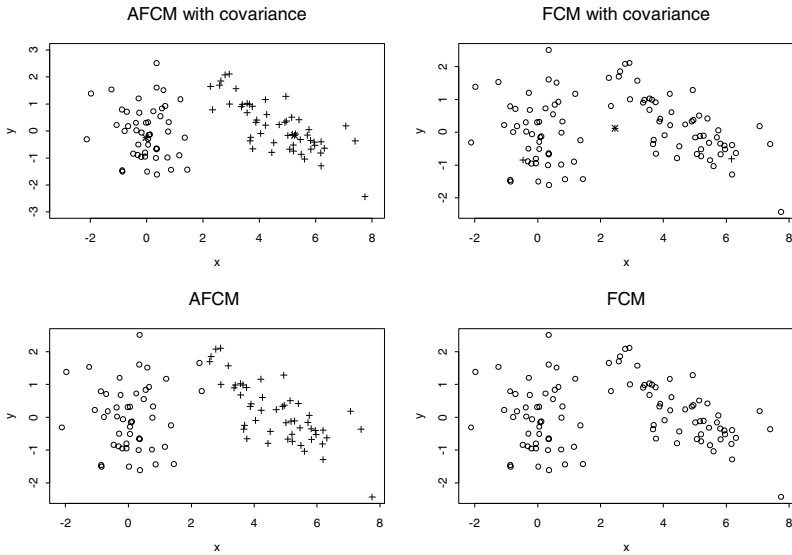


Fig. 3. Clustering results of FCM, AFCM, FCM-cov and AFCM-cov

clustering result. Compared with Example 1, we see that these two clusters are more close than Example 3. However, AFCM-L1 also bears the correct clustering results. But, FCM-L1 fails.

Example 5. To see the effect of outliers on FCM-cov or AFCM-cov, we add an outlying point (200,0) to the data set in Example 3. We then run both FCM-cov and AFCM-cov with $c = 2$. The clustering results are shown in Fig. 3. Compared with the results of Example 3 without outliers, we find that the FCM and FCM-cov cluster this outlier to a single cluster and all the rest have been grouped together into a separate cluster. It means that this outlier has heavily disturbed the original clustering of the data set in Example 3 when we implemented the FCM and FCM-cov clustering algorithms. However, AFCM-cov gives a perfect clustering result. Thus, the performance of AFCM-cov is the best among FCM, FCM-cov and AFCM in a noisy environment.

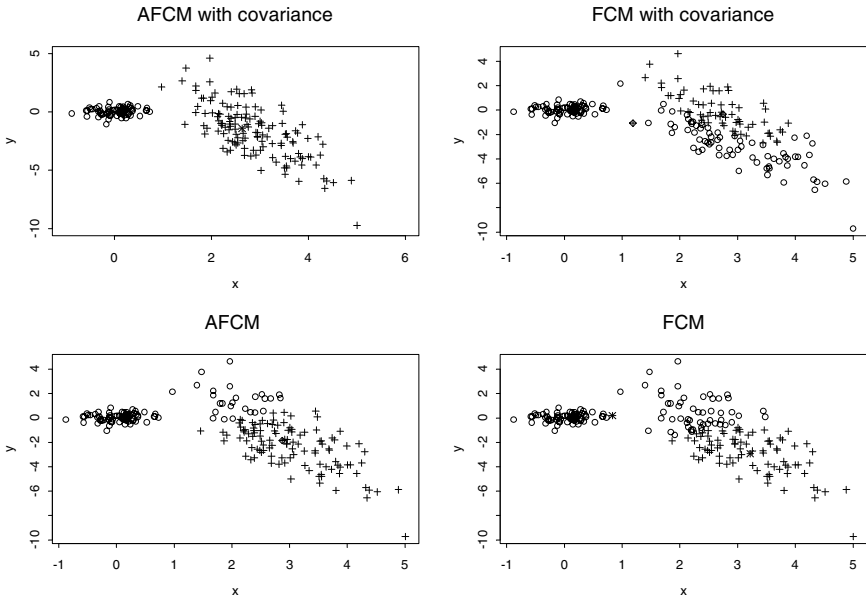


Fig. 4. Clustering results of FCM, AFCM, FCM-cov and AFCM-cov

Example 6. In this example, we consider a well-known clustering problem where there is an inordinate difference in the number of members in each sample cluster. The data set (sample size is 100) is generated from a 2-variate normal mixture

$$0.3N(\mu_1, \Sigma_1) + 0.7N(\mu_2, \Sigma_2) \text{ with } \mu_1 = (0,0), \mu_2 = (5,0), \Sigma_1 = \begin{bmatrix} 0.5 & 0 \\ 0 & 1 \end{bmatrix} \text{ and } \Sigma_2 = \begin{bmatrix} 2 & -1 \\ -1 & 1 \end{bmatrix}.$$

We run FCM, AFCM, FCM-cov and AFCM-cov with $c = 2$.

The clustering results are shown in Fig. 4. According to Fig. 4, we find that several data points represented as “o” move toward cluster 2 represented as “+” for FCM,

FCM-cov and AFCM. It means that FCM, FCM-cov and AFCM have inaccurate clustering result when the data set includes large numbers of different cluster sample sizes. However, AFCM-cov classified these two clusters without any misclassified data.

5 Conclusions

Cluster analysis is an unsupervised approach to pattern recognition. The FCM is the most used clustering algorithm. Because real data varies considerably, it is impossible for a clustering algorithm to fit all real cases. Therefore, there are many generalized types of FCM algorithms that have been recently unified into a generalized model. In this paper, we proposed AFCM- L_1 and AFCM-cov clustering algorithms to have better fitting to varieties of data sets. According to comparisons, the proposed algorithms have better accuracy and effectiveness than the existing methods.

References

1. Bezdek, J.C.: Pattern Recognition with Fuzzy Objective Function Algorithms. Plenum Press, New York (1981).
2. Bobrowski, L., Bezdek, J.C.: C-Means Clustering with the L_1 and L_∞ Norms. IEEE Trans. Systems, Man, and Cybernetics 21 (1991) 545-554.
3. Gustafson, D.E., Kessel, W.C.: Fuzzy Clustering with a Fuzzy Covariance Matrix. IEEE CDC, San Diego, CA (1979) 10-12.
4. Hathaway, R.J., Bezdek, J.C., Hu, Y.: Generalized Fuzzy C-Means Clustering Strategies Using L_p Norm Distances. IEEE Trans. Fuzzy Systems 8 (2000) 576-582.
5. Hoppner, F., Klawonn, F., Kruse, R., Runkler, T.: Fuzzy Cluster Analysis: Methods for Classification Data Analysis and Image Recognition. Wiley, New York (1999).
6. Jajuga, K.: L_1 -norm based fuzzy clustering, Fuzzy set and Systems 39 (1991) 43-50.
7. Kersten, P.R.: Fuzzy Order Statistics and Their Application to Fuzzy Clustering. IEEE Trans. Fuzzy Systems 7 (1999) 708-712.
8. Krishnapuram, R., Kim, J.: A Note on the Gustafson-Kessel and Adaptive Fuzzy Clustering Algorithms. IEEE Trans. Fuzzy Systems 7 (1999) 453-461.
9. Wu, K.L., Yang, M.S.: Alternative C-Means Clustering Algorithms. Pattern Recognition 35 (2002) 2267-2278.
10. Yang, M.S.: A Survey of Fuzzy Clustering. Mathematical and Computer Modeling 18 (1993) 1-16.
11. Yu, J., Yang, M.S.: Optimality Test for Generalized FCM and Its Application to Parameter Selection. IEEE Trans. Fuzzy Systems 13 (2005) 164-176.
12. Zadeh, L.A.: Fuzzy Sets. Information and Control 8 (1965) 338-353.

A Statistical Approach for Ownership Identification of Digital Images

Ching-Sheng Hsu¹, Shu-Fen Tu², and Young-Chang Hou³

¹ Department of Information Management, Ming Chuan University
5 De-Ming Rd., Gui Shan Township, Taoyuan County 333, Taiwan, R.O.C.
cshsu@mcu.edu.tw

² Department of Information Management, Chinese Culture University
No.55, Huagang Rd., Shihlin District, Taipei City 111, Taiwan, R.O.C.
dsf3@faculty.pccu.edu.tw

³ Department of Information Management, Tamkang University
151 Ying-Chuan Road, Tamsui, Taipei County 251, Taiwan, R.O.C.
ychou@mail.im.tku.edu.tw

Abstract. In this paper, we propose an ownership identification scheme for digital images with binary and gray-level ownership statements. The proposed method uses the theories and properties of sampling distribution of means to satisfy the requirements of robustness and security. Essentially, our method will not really insert the ownership statement into the host image. Instead, the ownership share will be generated by the sampling method as a key to reveal the ownership statement. Besides, our method allows ownership statements to be of any size and avoids the hidden ownership statement to be destroyed by the latter ones. When the rightful ownership of the image needs to be identified, our method can reveal the ownership statement without resorting to the original image. Finally, several common attacks to the image will be held to verify the robustness and the security is also analyzed.

1 Introduction

In recent years, more and more digital data such as image, audio, and video are transmitted and exchanged via Internet. However, in the cyberspace, the availability of duplication methods encourages the violation of intellectual property rights of digital data. Therefore, the protection of rightful ownership of digital data has become an important issue. Nowadays, many techniques have been developed to protect the rightful ownership of digital images. Digital watermarking, a kind of such techniques, is a method that hides a meaningful signature, or the so-called digital watermark, in an image for the purpose of copyright protection, integrity checking, and captioning. When the rightful ownership of the image needs to be identified, the hidden watermark can be extracted for the ownership verification. Digital watermarks can be either visible [1] or invisible [2–4]. In this paper, we shall focus on the invisible watermarks. In general, an effective watermarking scheme should satisfy certain requirements including imperceptibility, robustness, unambiguousness, security,

capacity, and low computational complexity [2, 5]. Some of these requirements may conflict each other and thereby introducing many technical challenges. Therefore, a reasonable compromise is required to achieve better performance for the intended applications.

Based on the taxonomy found in many literatures, we can group watermarking techniques into two categories: one is the spatial-domain approach [3, 4, 6], and the other is the transform-domain approach [2, 7–8]. During the watermark detection process, the original image may or may not be used. As the availability and portability are considered, those techniques that can reveal watermarks without resorting to the original image are preferred. Usually, the data of the host image should be adequately adjusted or altered for embedding the digital signature. Most related techniques use many pixels or transform coefficients to conceal one bit of information. Thus, the watermark should be much smaller than the host image so that the requirements of imperceptibility and robustness can be satisfied. Such property makes it impossible to embed a larger watermark into a smaller host image. Besides, if multiple watermarks need to be registered for a single digital image, it is also impossible for such methods to embed the latter watermark without destroying the former ones.

Recently, Chang et al. [7] proposed a copyright protection scheme which utilized visual cryptography concept and discrete cosine transformation (DCT) to satisfy the requirement of security and robustness. In their research, the DC coefficients of all DCT blocks are first extracted from the host image to form a master share; then an ownership share obtained by combining the master share and the watermark is constructed as a key to reveal the watermark without resorting to the original image. Their method requires the size of the watermark to be much smaller than that of the host image. For example, if the size of the original image is $M_1 \times M_2$, then the size of watermarks should be at most $M_1/92 \times M_2/92$ for gray-level and 256 colors. Therefore, it is quite impractical for their method to deal with gray-level watermarks. This method is quite different from the traditional watermarking schemes since nothing is inserted into the host image. Thus, we shall call the hidden images “ownership statements” instead of “watermarks” in the following.

In this paper, a copyright protection scheme without restricting the size of ownership statements is proposed. Our method does not need the image to be transformed between the spatial and frequency domains. Instead, the theories and properties of sampling distribution of means (SDM) are used to satisfy the requirements of security and robustness. This scheme has all the advantages of Chang’s method. For example, it does not need to alter the original image and can identify the ownership without resorting to the original image. Multiple ownership statements are allowed to be registered for a single image without causing any damage to other hidden ownership statements. In addition, it allows ownership statements to be of any size regardless of the size of the host image. Finally, we will prove that the proposed scheme is secure. Altogether, our method has more applications than copyright protection. For example, it can be applied to cover the transmission of confidential images.

2 Sampling Distribution of Means

According to the theory of sampling distribution in statistics, the sampling distribution of means (SDM) from a set of normally distributed data is also a normal distribution [9]. The arithmetic mean from a normally distributed population has several important mathematical properties, such as unbiasedness, efficiency, and consistency. The unbiased property says that the average of all the possible sample means of a given sample size n will be equal to the population mean μ . Note that the above properties are based on the assumption that the population itself is normally distributed. However, in many cases, the distribution of a population is unknown. In this case, the central limit theorem can be employed. According to the central limit theorem, as the sample size gets large enough, the sampling distribution of means can be approximated by the normal distribution. Statisticians have found a general rule that, for many population distributions, once the sample size is at least 30, the sampling distribution of the mean will be approximately normal.

Let $-\infty < X < +\infty$ be a normal random variable, denoted by $X \sim N(\mu, \sigma^2)$, which has a population mean μ and a standard deviation σ . We are able to transform all the observations of X to a new set of observations of a normal random variable Z with zero mean and standard deviation 1. This can be done by means of the transformation

$$Z = \frac{X - \mu}{\sigma}. \quad (1)$$

If samples of size n are drawn randomly from a population that has a mean of μ and a standard deviation of σ , the sample mean \bar{X} , $-\infty < \bar{X} < +\infty$, are approximately normally distributed for sufficiently large sample sizes regardless of the shape of the population distribution. From mathematical expectation, it can be shown that the mean of the sample means is the population mean:

$$\mu_{\bar{X}} = \mu \quad (2)$$

and the standard error of sample means is the standard deviation of the population divided by the square root of the sample size:

$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} \quad (3)$$

By the same token, \bar{X} can also be standardized to

$$Z = \frac{\bar{X} - \mu_{\bar{X}}}{\sigma_{\bar{X}}} = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \quad (4)$$

Theoretically, normal distribution is bell-shaped and symmetrical in its appearance, and the probability density function for Z is given by

$$\phi(Z) = \frac{1}{\sqrt{2\pi}} e^{-Z^2/2}. \quad (5)$$

Thus, for a fixed z , the probability of $Z \leq z$, denoted by $\Pr(Z \leq z) = \alpha$, can be computed by

$$\alpha = \int_{-\infty}^z \frac{1}{\sqrt{2\pi}} e^{-z^2/2} dz. \quad (6)$$

Inversely, for a given probability α , we define $Inv_NCD(\alpha)$ to be an *inverse normal cumulative density function* that can be used to find the corresponding z that satisfies Eq.(6). Therefore, using $Inv_NCD(\alpha)$, we can easily find $\lambda - 1$ partition points $z_r = Inv_NCD(r/\lambda)$ for $r = 1, 2, \dots, \lambda - 1$, which can partition the Z scale into λ equal-probability segments. For the extreme case $\lambda = 2$, $z_1 = 0$ can be used to partition the Z scale into two equal-probability segments $Z < 0$ and $Z \geq 0$. For further understanding the inverse normal cumulative density function $Inv_NCD(\alpha)$, interested readers may refer to [10–11].

3 The Proposed Scheme

3.1 Ownership Registration Phase

Assume that a copyright owner wants to cast a gray-level ownership statement W with $N_1 \times N_2$ pixels into a gray-level host image of any size for protecting his/her ownership. We also assume that the number of gray-levels is 256. Before we start to construct the ownership share O , the population mean μ and standard deviation σ of the pixel values of the host image should be calculated in advance. Besides, a random key L is used to generate the location of a pixel in the host image for sampling. For example, the first n elements are used to compute the first sample mean, the next n elements are used to compute the second sample mean, etc. Then, according to the central limit theorem and the unbiased property of SDM, we can form a normal distribution with the random variable \bar{X} which has the mean of μ and the standard error of σ/\sqrt{n} by sampling from the host image if the sample size is large enough. Then, the ownership share is generated by the following algorithm:

Algorithm Ownership Share Construction Procedure

- Input.* A gray-level host image H with any size, a gray-level ownership statement W with $N_1 \times N_2$ pixels, and a random key L .
- Output.* A gray-level ownership share O of size $N_1 \times N_2$ pixels.
- Step 1.* Compute the population mean μ and the standard deviation σ of the pixel values of the host image H .
- Step 2.* Generate 255 partition points z_1, z_2, \dots, z_{255} by $z_r = Inv_NCD(r/256)$ for $r = 1, 2, \dots, 255$. Then, the partition points are used to partition the Z scale into 256 equal-probability segments numbered from 0 to 255.
- Step 3.* Randomly select $n \geq 30$ pixel values x_1, x_2, \dots, x_n from the host image H (according to L) to form a sample mean \bar{x} , and then standardize the sample mean \bar{x} to $z = (\bar{x} - \mu)/(\sigma/\sqrt{n})$.

- Step 4.* If z falls in the segment g of the Z scale where $g \in \{0, 1, \dots, 255\}$, then the pixel gray-level $m_{i,j}$ of the master share M is $m_{i,j} = g$.
- Step 5.* For the pixel gray-level $w_{i,j}$ of the ownership statement W , determine the corresponding pixel gray-level $o_{i,j}$ of the ownership share O according to $o_{i,j} = w_{i,j} \oplus m_{i,j}$, where \oplus denotes the bit-wise XOR operation.
- Step 6.* Repeat Step 3 to Step 5 until all pixels of the ownership statement are processed.

Finally, the random key L must be kept secretly by the copyright owner, and the ownership share O should be registered with a trusted third party for further authentication.

3.2 Ownership Identification Phase

In the ownership identification phase, the copyright owner should provide the same secret key L used in the ownership registration phase so that the correct sequence of pixel locations can be obtained during the sampling process. Then, the hidden ownership statement is recovered by the following algorithm:

Algorithm *Ownership statement Revelation Procedure*

- Input.* A gray-level host image H' with any size, a gray-level ownership share O with $N_1 \times N_2$ pixels, and a secret key L .
- Output.* A recovered ownership statement W' of size $N_1 \times N_2$ pixels.
- Step 1.* Compute the population mean μ' and the standard deviation σ' of the pixel values of the host image H' .
- Step 2.* Generate 255 partition points z_1, z_2, \dots, z_{255} by $z_r = \text{Inv_NCD}(r/256)$ for $r = 1, 2, \dots, 255$. Then, the partition points are used to partition the Z scale into 256 equal-probability segments numbered from 0 to 255.
- Step 3.* Randomly select $n \geq 30$ pixel values x'_1, x'_2, \dots, x'_n from the host image H' (according to L) to form a sample mean \bar{x}' , and then standardize the sample mean \bar{x}' to $z = (\bar{x}' - \mu') / (\sigma' / \sqrt{n})$.
- Step 4.* If z falls in the segment g of the Z scale where $g \in \{0, 1, \dots, 255\}$, then the pixel gray-level $m'_{i,j}$ of the master share M is $m'_{i,j} = g$.
- Step 5.* For the pixel gray-level $o_{i,j}$ of the ownership share O , determine the corresponding pixel gray-level $w'_{i,j}$ of the ownership statement W' according to $w'_{i,j} = o_{i,j} \oplus m'_{i,j}$, where \oplus denotes the bit-wise XOR operation.
- Step 6.* Repeat Step 3 to Step 5 until all pixels of the ownership share are processed.

Note that the controversial image H' may be altered or modified by the image processing filters or compression techniques. Consequently, the recovered ownership statement W' may be different from the original ownership statement W to some extent.

4 Results and Analysis

In this section, the robustness of our scheme against several common attacks is examined. Besides, the security is also analyzed. In the following experiments, the sample size $n = 30$ is used to proceed the sampling process. The gray-level host image with 512×512 pixels is shown in Fig. 1(a), and the binary and gray-level ownership statements with 400×400 pixels are shown in Fig. 1(b) and 1(c), respectively. Besides, the corresponding binary and gray-level ownership shares, generated from the original host image (Fig. 1(a)), are shown in Fig. 2(a) and 2(b), respectively. The similarity between two binary images is measured by the normalized correlation (NC) and that between two gray-level images is measured by the peak signal-to-noise ratio (PSNR).

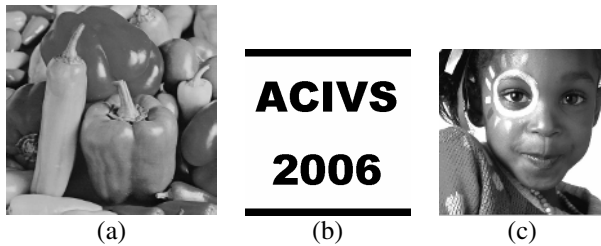


Fig. 1. (a) The gray-level host image (512×512 pixels); (b) the binary ownership statement (400×400 pixels); (c) the gray-level ownership statement (400×400 pixels)

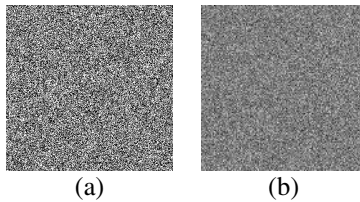


Fig. 2. (a) The binary ownership share; (b) the gray-level ownership share (400×400 pixels)

4.1 Security Analysis

According to the central limit theorem and the unbiased property of SDM, we have that $\Pr(\bar{x} \geq \mu) = \Pr(\bar{x} < \mu) = 0.5$ if the sample size n is large enough. Thus, the ratios of black and white pixels on the binary master share will be approximately 50% to 50%. In other words, we can expect that the probability of a pixel on the binary master share is black will be 0.5. Therefore, on the ownership share, the expected ratio of black pixels is also 0.5. In Fig. 3, we show the ratios of black pixels of ownership shares for several well-known images under the sample size $n = 30$. The result shows that the ratios of black and white pixels of the ownership shares are approximately 50% to 50%. Fig. 4 shows the histograms of the gray-level ownership shares under the sample size $n = 200$. As we can see from Fig. 4, the ratios of all of the gray-levels are nearly the same. Thus, we can conclude that the central limit theorem and the unbiased property of SDM hold and the security of the proposed scheme is ensured.

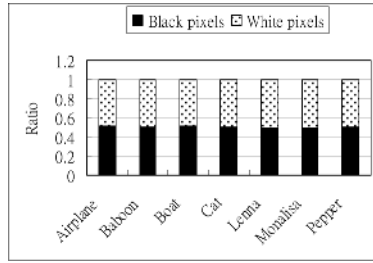


Fig. 3. Ratios of black and white pixels of the ownership shares

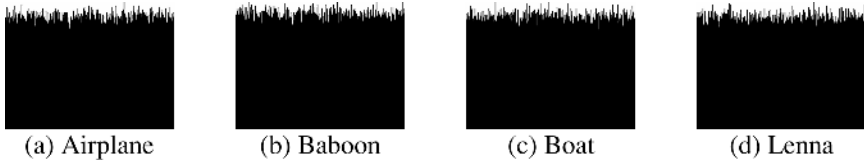


Fig. 4. Histograms of the gray-level ownership shares

4.2 Robustness Analysis

Several experiments are performed to demonstrate the robustness of the proposed scheme against several common attacks, including JPEG lossy compression, sharpening, lightening, darkening, noising, cropping, blurring, distorting, rescaling, and jitter attacks. Note that the noised image is with 10% monochromatic noises. The compression ratio of the JPEG attack is 5:1. The cropped attack is to erase the top left area (about $1/3 \times 1/3$) of the image. The rescaled image is obtained by first downscaling the image by a factor of 2 in each direction and then upscaling the downscaled image to the original size. Besides, the jitter attack is used to remove two distinct columns (with the width of five pixels each) on the left half of the image and then insert them into the other positions on the right half.

Table 1. Robustness analysis of the proposed scheme

Attacks	PSNR ^(H) (dB)	NC (%)	PSNR ^(W) (dB)
JPEG (compression ratio = 5:1)	38.90	98.18	20.66
Sharpening	26.42	94.78	16.31
Lightening	18.59	100.0	$+\infty$
Darkening	18.59	98.63	21.85
10% noising	24.44	90.92	14.00
11% cropping	18.84	76.1	10.68
Blurring	26.71	93.96	15.63
Distorting	21.93	88.33	13.19
Rescaling	32.91	97.71	19.58
Jitter	20.56	83.51	11.92

In Table 1, $PSNR^{(H)}$ is the similarity of the attacked host image, and NC and $PSNR^{(W)}$ are the similarities of the recovered binary and gray-level ownership statements, respectively. According to the experimental results shown in Table 1, Fig.5, and Fig. 6, we can find that JPEG, sharpening, lightening, darkening, rescaling, and blurring attacks cause little damage to the revealed ownership statements. Although some of the attacks, such as lightening, darkening, cropping, distorting, and jitter attacks, may lead to low $PSNR^{(H)}$ values of the attacked images, it seems that the recovered ownership statements can also be clearly identified by human eyes. Besides, it can be seen from the results that the proposed method can effectively resist the lightening and darkening attacks. Totally speaking, we can conclude that our scheme meets the requirements of unambiguousness and robustness against several common attacks.

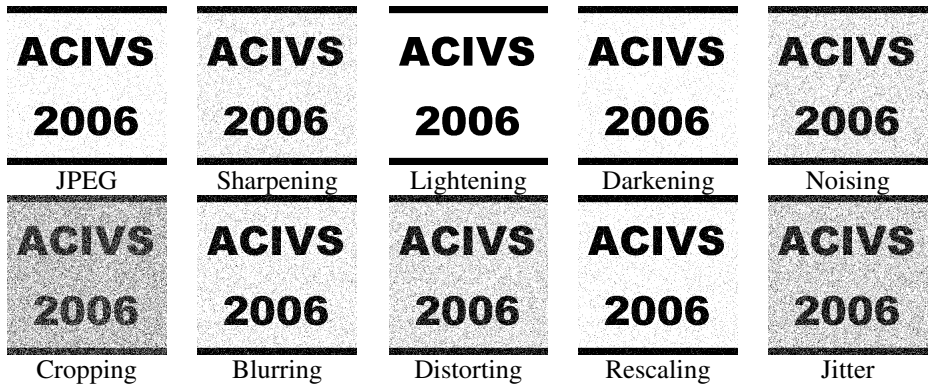


Fig. 5. The recovered binary ownership statements (400×400 pixels) upon different attacks

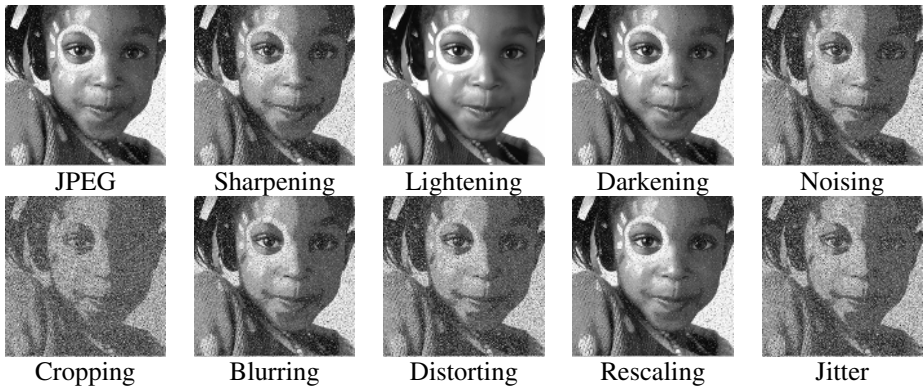


Fig. 6. The recovered gray-level ownership statements (400×400 pixels) upon different attacks

5 Conclusions

In this paper, a method that can use binary and gray-level ownership statements to protect the rightful ownership of digital images was proposed. Our scheme employed

the sampling distribution of means to satisfy the requirements of robustness and security. Based on the sampling method, it is possible to deal with binary and gray-level ownership statements of any size. Since the scheme does not really insert the ownership statement into the image to be protected, the image will not be altered and the rightful ownership can be identified without resorting to the original image. Moreover, it also allows multiple ownership statements to be cast into a single host image without causing any damage to other hidden ownership statements. As shown in the experimental results, we could conclude that the proposed scheme can resist several common attacks. Besides, without the correct secret key, no one can recover any meaningful ownership statements from the host image. Thus, the proposed scheme is secure. Since security is ensured, the proposed scheme is also suitable to cover the transmission of secret images. In the future, the issue of color ownership statements will be studied.

Acknowledgement

This work was supported in part by a grant from National Science Council of the Republic of China under the project NSC-93-2213-E-032-033.

References

1. Braudaway, G.W., Magerlein, K.A., and Mintzer, F.: Protecting Publicly-available Images with a Visible Image Watermark. *Proc. SPIE*. Vol. 2659 (1996) 126–133
2. Cox, I.J., Kilian, J., Leighton, T., and Shamoon, T.: Secure spread spectrum watermarking for multimedia. *IEEE Trans. Image Process.* Vol. 6, No. 12 (1997) 1673–1687
3. Low, S. and Maxemchuk, N.: Performance Comparison of Two Text Marking Methods. *IEEE J. Selected Areas in Communications*. Vol. 16, No. 4 (1998) 561–572
4. Ohbuchi, R., Masuda, H., and Aono, M.: Watermarking Three-Dimensional Polygonal Models through Geometric and Topological Modifications. *IEEE J. Selected Areas in Communications*. Vol. 16, No. 4 (1998) 551–560
5. Katzenbeisser, S. and Petitcolas, F.A.P.: *Information Hiding Techniques for Steganography and Digital Watermarking*. Artech house, Norwood, MA (2000) 101–109
6. Hou, Y.C. and Chen, P.M.: An Asymmetric Watermarking Scheme Based on Visual Cryptography. *Proc. Fifth Signal Process. Conf.* Vol. 2 (2000) 992–995
7. Chang, C.C., Hsiao, J.Y., and Yeh, J.C.: A Colour Image Copyright Protection Scheme Based on Visual Cryptography and Discrete Cosine Transform. *The Imaging Sci. J.*, Vol. 50, (2002) 133–140
8. Kim, W.S., Hyung, O.H., and Park, R.H.: Wavelet Based Watermarking Method for Digital Images using the Human Visual System. *Electron. Lett.* Vol. 35 (1999) 466–468
9. Berenson, M.L. and Levine, D.M.: *Basic Business Statistics: Concepts and Applications*. Prentice-Hall, New Jersey (1999) 337–353
10. Acklam, P.J.: An Algorithm for Computing the Inverse Normal Cumulative Distribution Function (2004). Available: <http://home.online.no/~pjacklam/notes/invnorm/>
11. Bialas, W.F.: *Lecture Notes in Applied Probability*. Department of Industrial Engineering, the State University of New York at Buffalo (Summer 2004)

Rigid and Non-rigid Face Motion Tracking by Aligning Texture Maps and Stereo-Based 3D Models*

Fadi Dornaika and Angel D. Sappa

Computer Vision Center
Campus UAB
08193 Bellaterra, Barcelona, Spain
{dornaika, sappa}@cvc.uab.es

Abstract. Accurate rigid and non-rigid tracking of faces is a challenging task in computer vision. Recently, appearance-based 3D face tracking methods have been proposed. These methods can successfully tackle the image variability and drift problems. However, they may fail to provide accurate out-of-plane face motions since they are not very sensitive to out-of-plane motion variations. In this paper, we present a framework for fast and accurate 3D face and facial action tracking. Our proposed framework retains the strengths of both appearance and 3D data-based trackers. We combine an adaptive appearance model with an on-line stereo-based 3D model. We provide experiments and performance evaluation which show the feasibility and usefulness of the proposed approach.

1 Introduction

The ability to detect and track human heads and faces in video sequences is useful in a great number of applications, such as human-computer interaction and gesture recognition. There are several commercial products capable of accurate and reliable 3D head position and orientation estimation (e.g., the acoustic tracker system Mouse [www.vrdepot.com/vrteclg.htm]). These are either based on magnetic sensors or on special markers placed on the face; both practices are encumbering, causing discomfort and limiting natural motion. Vision-based 3D head tracking provides an attractive alternative since vision sensors are not invasive and hence natural motions can be achieved [1]. However, detecting and tracking faces in video sequences is a challenging task due to the image variability caused by pose, expression, and illumination changes.

Recently, deterministic and statistical appearance-based 3D head tracking methods have been proposed and used by some researchers [2, 3, 4]. These methods can successfully tackle the image variability and drift problems by using deterministic or statistical models for the global appearance of a special object class: the face. However, appearance-based methods dedicated to full 3D head tracking may suffer from some inaccuracies since these methods are not very sensitive to out-of-plane motion variations. On the other hand, the use of dense 3D facial data provided by a stereo rig or a

* This work was supported by the MEC project TIN2005-09026 and The Ramón y Cajal Program.

range sensor can provide very accurate 3D face motions. However, computing the 3D face motions from the stream of dense 3D facial data is not straightforward. Indeed, inferring the 3D face motion from the dense 3D data needs an additional process. This process can be the detection of some particular facial features in the range data/images from which the 3D head pose can be inferred. For example, in [5], the 3D nose ridge is detected and then used for computing the 3D head pose. Alternatively, one can perform a registration between 3D data obtained at different time instants in order to infer the relative 3D motions. The most common registration technique is the Iterative Closest Point (ICP) [6] algorithm. This algorithm and its variants can provide accurate 3D motions but their significant computational cost prohibits real-time performance.

The main contribution of this paper is a robust 3D face tracker that combines the advantages of both appearance-based trackers and 3D data-based trackers while keeping the CPU time very close to that required by real-time trackers. In our work, we use the deformable 3D model *Candide* [7] which is a simple model embedding non-rigid facial motion using the concept of facial actions. Our proposed framework for tracking faces in videos can be summarized as follows. First, the 3D head pose and some facial actions are estimated from the monocular image by registering the warped input texture with a shape-free facial texture map. Second, based on these current parameters the 2D locations of the mesh vertices are inferred by projecting the current mesh onto the current video frame. Then the 3D coordinates of these vertices are computed by stereo reconstruction. Third, the relative 3D face motion is then obtained using a robust 3D-to-3D registration technique between two meshes corresponding to the first video frame and the current video frame, respectively. Our framework attempts to reduce the number of outlier vertices by deforming the meshes according to the same current facial actions and by exploiting the symmetrical shape of the 3D mesh.

The resulting 3D face and facial action tracker is accurate, fast, and drift insensitive. Moreover, unlike many proposed frameworks (e.g., [8]), it does not require any learning stage since it is based on online facial appearances and online stereo 3D data.

The remainder of the paper proceeds as follows. Section 2 introduces our deformable 3D facial model. Section 3 states the problem we are focusing on, and describes the online adaptive appearance model. Section 4 summarizes the appearance-based monocular tracker that tracks in real-time the 3D head pose and some facial actions. It gives some evaluation results. Section 5 describes a robust 3D-to-3D registration that combines the monocular tracker's results and the stereo-based reconstructed vertices. Section 6 gives some experimental results.

2 Modeling Faces

In this section, we briefly describe our deformable face model and explain how to produce a shape-free facial texture map.

A Deformable 3D Model. As mentioned before, we use the 3D face model *Candide*. This 3D deformable wireframe model was first developed for the purpose of model-based image coding and computer animation. The 3D shape of this wireframe model is directly recorded in coordinate form. It is given by the coordinates of the 3D vertices

$\mathbf{P}_i, i = 1, \dots, n$ where n is the number of vertices. Thus, the shape up to a global scale can be fully described by the $3n$ -vector \mathbf{g} ; the concatenation of the 3D coordinates of all vertices \mathbf{P}_i . The vector \mathbf{g} is written as:

$$\mathbf{g} = \mathbf{g}_s + \mathbf{A} \boldsymbol{\tau}_a \quad (1)$$

where \mathbf{g}_s is the static shape of the model, $\boldsymbol{\tau}_a$ the animation control vector, and the columns of \mathbf{A} are the Animation Units. In this study, we use six modes for the facial Animation Units (AUs) matrix \mathbf{A} . Without loss of generality, we have chosen the six following AUs: lower lip depressor, lip stretcher, lip corner depressor, upper lip raiser, eyebrow lowerer and outer eyebrow raiser. These AUs are enough to cover most common facial animations (mouth and eyebrow movements). Moreover, they are essential for conveying emotions.

In equation (1), the 3D shape is expressed in a local coordinate system. However, one should relate the 3D coordinates to the image coordinate system. To this end, we adopt the weak perspective projection model. We neglect the perspective effects since the depth variation of the face can be considered as small compared to its absolute depth. Thus, the state of the 3D wireframe model is given by the 3D head pose parameters (three rotations and three translations) and the internal face animation control vector $\boldsymbol{\tau}_a$. This is given by the 12-dimensional vector \mathbf{b} :

$$\mathbf{b} = [\theta_x, \theta_y, \theta_z, t_x, t_y, t_z, \boldsymbol{\tau}_a^T]^T \quad (2)$$

Shape-free Facial Texture Maps. A face texture is represented as a shape-free texture (geometrically normalized image). The geometry of this image is obtained by projecting the static shape \mathbf{g}_s using a centered frontal 3D pose onto an image with a given resolution. The texture of this geometrically normalized image is obtained by texture mapping from the triangular 2D mesh in the input image (see figure 1) using a piece-wise affine transform, \mathcal{W} . The warping process applied to an input image \mathbf{y} is denoted by:

$$\mathbf{x}(\mathbf{b}) = \mathcal{W}(\mathbf{y}, \mathbf{b}) \quad (3)$$

where \mathbf{x} denotes the shape-free texture map and \mathbf{b} denotes the geometrical parameters. Several resolution levels can be chosen for the shape-free textures. The reported results are obtained with a shape-free patch of 5392 pixels.

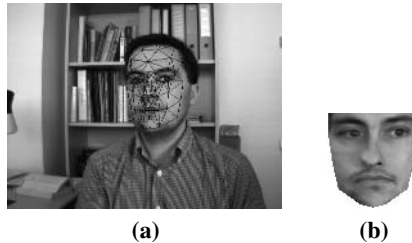


Fig. 1. (a) an input image with correct adaptation. (b) the corresponding shape-free facial map.

3 Problem Formulation and Facial Texture Model

Given a monocular video sequence depicting a moving head/face, we would like to recover, for each frame, the 3D head pose and the facial actions encoded by the control vector τ_a . In other words, we would like to estimate the vector \mathbf{b}_t (equation 2) at time t given all the observed data until time t , denoted $\mathbf{y}_{1:t} \equiv \{\mathbf{y}_1, \dots, \mathbf{y}_t\}$. In a tracking context, the model parameters associated with the current frame will be handed over to the next frame. For each input frame \mathbf{y}_t , the observation is simply the shape-free texture map associated with the geometric parameters \mathbf{b}_t . We use the HAT symbol for the tracked parameters and textures. For a given frame t , $\hat{\mathbf{b}}_t$ represents the computed geometric parameters and $\hat{\mathbf{x}}_t$ the corresponding shape-free texture map, that is,

$$\hat{\mathbf{x}}_t = \mathbf{x}(\hat{\mathbf{b}}_t) = \mathcal{W}(\mathbf{y}_t, \hat{\mathbf{b}}_t) \tag{4}$$

The estimation of $\hat{\mathbf{b}}_t$ from the sequence of images will be presented in the next Section.

By assuming that the pixels within the shape-free patch are independent, we can model the facial appearance using a multivariate Gaussian with a diagonal covariance matrix Σ . The choice of a Gaussian distribution is motivated by the fact that this kind of distribution provides simple and general model for additive noises. In other words, this multivariate Gaussian is the distribution of the facial texture maps $\hat{\mathbf{x}}_t$. Let $\boldsymbol{\mu}$ be the Gaussian center and $\boldsymbol{\sigma}$ the vector containing the square root of the diagonal elements of the covariance matrix Σ . $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are d -vectors (d is the size of \mathbf{x}). Although the independence assumption may be violated, at least locally, we adopt it in our work in order to keep the problem tractable. In summary, the observation likelihood at time t is written as

$$p(\mathbf{y}_t | \mathbf{b}_t) = p(\mathbf{x}_t | \mathbf{b}_t) = \prod_{i=1}^d \mathbf{N}(x_i; \mu_i, \sigma_i)_t \tag{5}$$

where $\mathbf{N}(x_i; \mu_i, \sigma_i)$ is a normal density:

$$\mathbf{N}(x_i; \mu_i, \sigma_i) = (2\pi\sigma_i^2)^{-1/2} \exp \left[-\rho \left(\frac{x_i - \mu_i}{\sigma_i} \right) \right], \quad \rho(x) = \frac{1}{2} x^2 \tag{6}$$

We assume that the appearance model summarizes the past observations under an exponential envelope, that is, the past observations are exponentially forgotten with respect to the current texture. When the appearance is tracked for the current input image, *i.e.* the texture $\hat{\mathbf{x}}_t$ is available, we can compute the updated appearance and use it to track in the next frame.

When the appearance is tracked for the current input image, *i.e.* the texture $\hat{\mathbf{x}}_t$ is available, we can compute the updated appearance and use it to track in the next frame.

It can be shown that the appearance model parameters, *i.e.*, the μ_i 's and σ_i 's can be updated from time t to time $(t + 1)$ using the following equations (see [9] for more details on OAMs):

$$\mu_{i(t+1)} = (1 - \alpha) \mu_{i(t)} + \alpha \hat{x}_{i(t)} \tag{7}$$

$$\sigma_{i(t+1)}^2 = (1 - \alpha) \sigma_{i(t)}^2 + \alpha (\hat{x}_{i(t)} - \mu_{i(t)})^2 \tag{8}$$

This technique, also called recursive filtering, is simple, time-efficient and therefore, suitable for real-time applications. The appearance parameters reflect the most recent observations within a roughly $L = 1/\alpha$ window with exponential decay. Note that $\boldsymbol{\mu}$ is initialized with the first patch $\hat{\mathbf{x}}_0$. In order to get stable values for the variances, equation (8) is not used until the number of frames reaches a given value (e.g., the first 40 frames). For these frames, the classical variance is used, that is, equation (8) is used with α being set to $\frac{1}{t}$.

4 Tracking by Aligning Facial Texture Maps

We consider the state vector $\mathbf{b} = [\theta_x, \theta_y, \theta_z, t_x, t_y, t_z, \boldsymbol{\tau}_a^T]^T$ encapsulating the 3D head pose and the facial actions. In [10], we have developed a fast method to compute this state from the previous known state $\hat{\mathbf{b}}_{t-1}$ and the current input image \mathbf{y}_t . An overview of this method is presented here.

The sought geometrical parameters \mathbf{b}_t at time t are estimated using a region-based registration technique that does not need any image feature extraction. For this purpose, we minimize the *Mahalanobis* distance between the warped texture and the current appearance mean - the current Gaussian center $\boldsymbol{\mu}_t$

$$\min_{\mathbf{b}_t} D(\mathbf{x}(\mathbf{b}_t), \boldsymbol{\mu}_t) = \min_{\mathbf{b}_t} \sum_{i=1}^d \left(\frac{x_i - \mu_i}{\sigma_i} \right)^2 \quad (9)$$

The above criterion can be minimized using iterative first-order linear approximation which is equivalent to a Gauss-Newton method where the initial solution is given by the previous known state $\hat{\mathbf{b}}_{t-1}$. It is worthwhile noting that the minimization is equivalent to maximizing the likelihood measure given by (5). In the above optimization, the gradient matrix $\frac{\partial \mathcal{W}(\mathbf{y}_t, \mathbf{b}_t)}{\partial \mathbf{b}_t} = \frac{\partial \mathbf{x}_t}{\partial \mathbf{b}_t}$ is computed for each frame and is approximated by numerical differences similarly to the work of Cootes [11].

On a 3.2 GHz PC, a non-optimized C code of the approach computes the 3D head pose and the six facial actions in 50 ms. About half that time is required if one is only interested in computing the 3D head pose parameters.

Accuracy evaluation. In [12], we have evaluated the accuracy of the above proposed monocular tracker. To this end, we have used ground truth data that were recovered by the Iterative Closest Point algorithm [6] and dense 3D facial data. Figure 2 depicts the monocular tracker errors associated with a 300-frame long sequence which contains rotational and translational out-of-plane head motions. The nominal absolute depth of the head was about 65 cm, and the focal length of the camera was 824 pixels. As can be seen, the out-of-plane motion errors can be large for some frames for which there is a room for improvement. Moreover, this evaluation has confirmed the general trend of appearance-based trackers, that is, the out-of-plane motion parameters (pitch angle, yaw angle, and depth) are more affected by errors than the other parameters. We point out that the facial feature motions obtained by the above appearance-based tracker can be accurately recovered. Indeed, these features (the lips and the eyebrows) have specific textures, so their independent motion can be accurately recovered by the appearance-based tracker.

One expects that the monocular tracker accuracy can be improved if an additional cue is used. In our case, the additional cue will be the 3D data associated with the mesh vertices provided by stereo reconstruction. Although the use of stereo data may seem as an excess requirement, recall that cheap and compact stereo systems are now widely available (e.g., [www.ptgrey.com]). We point out that stereo data are used to refine the static model \mathbf{g}_s in the sense that the facial mesh can be more person-specific.

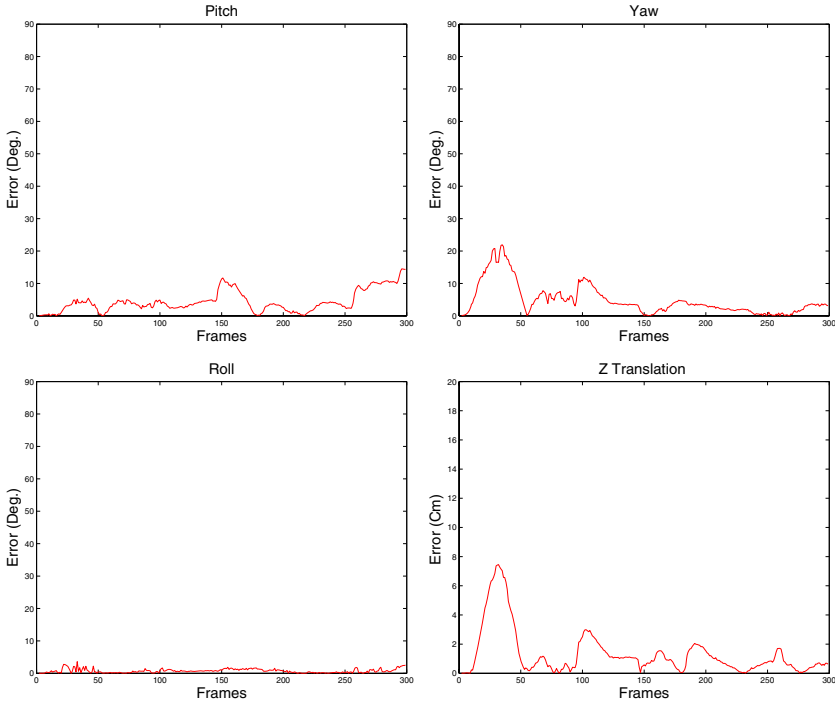


Fig. 2. 3D face motion errors computed by the ICP algorithm associated with a 300-frame long sequence

5 Tracking by Aligning Texture Maps and Stereo-Based 3D Models

In this section, we propose a novel tracking scheme that aims at computing a fast and accurate 3D face motion. To this end, we exploit the tracking results provided by the appearance-based tracker (Section 4) and the availability of a stereo system for reconstructing the mesh vertices. The whole algorithm is outlined in Figure 3. Note that the facial actions are already computed using the technique described in Section 4.

Our approach to 3D face tracking is simple and can be stated as follows: *If the 3D coordinates of the 3D mesh vertices at two different time instants are given in the same coordinate system, then the rigid transform corresponding to the 3D face motion can easily be recovered using a robust 3D point-to-point registration algorithm.* Without

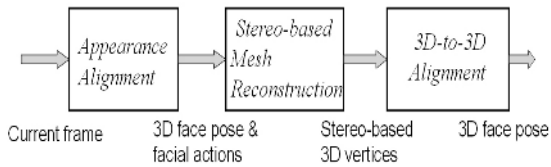


Fig. 3. The main steps of the developed robust 3D face tracker

loss of generality, the 3D face motion will be expressed with respect to the head coordinate system associated with the first video frame¹. In order to invoke the registration algorithm one has to compute the 3D coordinates of the vertices associated with the two different video frames: the initial video frame and the current one (see Figure 4). This is carried out using stereo-based data associated with the first frame and the current frame. Recall that the first 3D head pose can be inferred using a classical model-based pose estimation algorithm. The proposed algorithm can be summarized as follows.

1. Invoke the appearance-tracker to recover the current 3D head pose and facial actions (Section 4).
2. Based on the estimated 3D head pose and facial actions, deform the 3D mesh and project it onto the current frame.
3. Reconstruct the obtained image points (stereo reconstruction of the mesh vertices).
4. Deform the initial mesh (first frame) according to the current estimated facial actions.
5. Eliminate the vertices that are not consistent with the 3D symmetry test. Recall that the Euclidean distance between two symmetrical vertices is invariant due to the model symmetry.
6. Invoke a robust registration technique that provides the rigid displacement corresponding to the actual 3D face motion between the initial frame and the current frame. This step is detailed in Figure 5.

As can be seen, the recovered 3D face motion has relied on both the appearance model and the stereo-based 3D data. Steps 4 and 5 have been introduced in order to reduce the number of outlier vertices. Since in step 4, the initial mesh is deformed according to the current facial expression, a rigid registration technique can be efficiently applied. Recall that for a profile view the vertices associated with the hidden part of the face may have erroneous depth due to occlusion. Thus, step 5 eliminates the vertices with erroneous depth since they do not satisfy the symmetry constraint. Reducing the number of outliers is very useful for obtaining a very fast robust registration in the sense that a very few random samples are needed. Note that although the appearance-based tracker may provide slightly inaccurate out-of-plane parameters, the corresponding projected mesh onto the current image is still useful for getting the current stereo-based 3D coordinates of the mesh vertices (steps 2 and 3).

¹ Upgrading this relative 3D face motion to a 3D head pose that is expressed in the camera coordinate system is carried out using the 3D head pose associated with the first video frame that can be inferred using a classical 3D pose estimation algorithm.

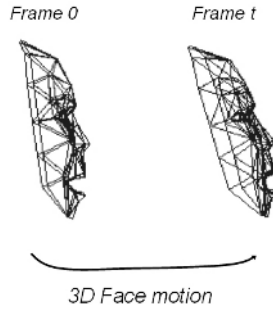


Fig. 4. The relative 3D face motion is recovered using a robust 3D-to-3D registration

We stress the fact that the proposed approach is not similar to a classical stereo-based 3D tracker where feature points are tracked across the image sequence. In our method, there is no feature matching and tracking across the image sequence. Instead, the whole face appearance is tracked in order to accurately locate the facial features (the projection of the mesh vertices) from which the 3D coordinates are inferred.

Robust 3D registration methods have been proposed in recent literature (e.g., see [13, 14]). In our work, we use a RANSAC-like technique that computes an adaptive threshold for inlier/outlier detection.

Inlier detection. The question now is: Given a subsample k and its associated solution \mathbf{D}_k , How do we decide whether or not an arbitrary vertex is an inlier? In techniques dealing with 2D geometrical features (points and lines) [15], this is achieved using the distance in the image plane between the actual location of the feature and its mapped location. If this distance is below a given threshold then this feature is considered as an inlier; otherwise, it is considered as an outlier. Here we can do the same by manually defining a distance in 3D space. However, this fixed selected threshold cannot accommodate all cases and all noises. Therefore, we use an adaptive threshold distance that is computed from the residual errors associated with all subsamples. Our idea is to compute a robust estimation of standard deviation of the residual errors. In the exploration step, for each subsample k , the median of residuals was computed. If we denote by \overline{M} the least median among all K medians, then a robust estimation of the standard deviation of the residuals is given by [16]:

$$\hat{\sigma} = 1.4826 \left[1 + \frac{5}{N-3} \right] \sqrt{\overline{M}} \quad (10)$$

where N is the number of vertices. Once $\hat{\sigma}$ is known, any vertex j can be considered as an inlier if its residual error satisfies $|r_j| < 3\hat{\sigma}$.

Computational cost. On a 3.2 GHz PC, a non-optimized C code of the robust 3D-to-3D registration takes about 10ms assuming that the number of random samples K is set to 8 and the total number of the 3D mesh vertices, N , is 113. This computational time includes both the stereo reconstruction and the robust technique outlined in Figure 5. Thus, by appending the robust 3D-to-3D registration to the appearance-based tracker (described before) a video frame can be processed in about 60 ms.

Random sampling: Repeat the following three steps K times

1. Draw a random subsample of 3 different pairs of vertices. We have three pairs of 3D points $\{\mathbf{M}_i \leftrightarrow \mathbf{S}_i\}$, $i = 1, 2, 3$. \mathbf{M}_i denotes the 3D coordinates of vertex i associated with the first frame, and \mathbf{S}_i denotes the 3D coordinates of the same vertex with the current frame t . \mathbf{M}_i and \mathbf{S}_i are expressed in the same coordinate system.
2. For this subsample, indexed by k ($k = 1, \dots, K$), compute the 3D rigid displacement $\mathbf{D}_k = [\mathbf{R}_k | \mathbf{T}_k]$, where \mathbf{R}_k is a 3D rotation and \mathbf{T}_k a 3D translation, that brings these three pairs into alignment. \mathbf{R}_k and \mathbf{T}_k are computed by minimizing the residual error $\sum_{i=1}^3 |\mathbf{S}_i - \mathbf{R}_k \mathbf{M}_i - \mathbf{T}_k|^2$. This is carried out using the quaternion method [17].
3. For this solution \mathbf{D}_k , compute the median M_k of the squared residual errors with respect to the whole set of N vertices. Note that we have N residuals corresponding to all vertices $\{\mathbf{M}_j \leftrightarrow \mathbf{S}_j\}$, $j = 1, \dots, N$. The squared residual associated with an arbitrary vertex \mathbf{M}_j is $|\mathbf{S}_j - \mathbf{R}_k \mathbf{M}_j - \mathbf{T}_k|^2$.

Solution:

1. For each solution $\mathbf{D}_k = [\mathbf{R}_k | \mathbf{T}_k]$, $k = 1, \dots, K$, compute the number of inliers among the entire set of vertices (see text). Let n_k be this number.
2. Choose the solution that has the largest number of inlier vertices.
3. Refine the corresponding solution using all its inlier pairs.

Fig. 5. Recovering the relative 3D face motion using online stereo and robust statistics

6 Experimental Results

We use the stereo system Bumblebee from Point Grey [www.ptgrey.com]. It consists of two Sony ICX084 color CCDs with 6mm focal length lenses. The monocular sequence is used by the appearance-tracker (Section 4), while the stereo sequence is used by the 3D-to-3D registration technique (Section 5). Figure 6 (Top) displays the face and facial action tracking results associated with a 300-frame-long sequence (only three frames are shown). The tracking results were obtained using the proposed framework described in Sections 4 and 5. The upper left corner of each image shows the current appearance ($\boldsymbol{\mu}_t$) and the current shape-free texture ($\hat{\mathbf{x}}_t$). In this sequence, the nominal absolute depth of the head was about 65 cm.

As can be seen, the tracking results indicate good alignment between the mesh model and the images. However, it is very difficult to evaluate the accuracy of the out-of-plane motions by only inspecting the projection of the 3D wireframe onto these 2D images. Therefore, we have run three different methods using the same video sequence. The first method is given by the appearance-based tracker (Section 4). The second method is given the proposed method (Sections 4 and 5). Note that the number of random samples used by the proposed method is set to 8. The third method is given by the ICP registration between dense facial surfaces where the facial surface model is set to the one obtained with the first stereo pair [12]. Since the ICP registration results are accurate we can use them as ground-truth data for the relative 3D face motions.

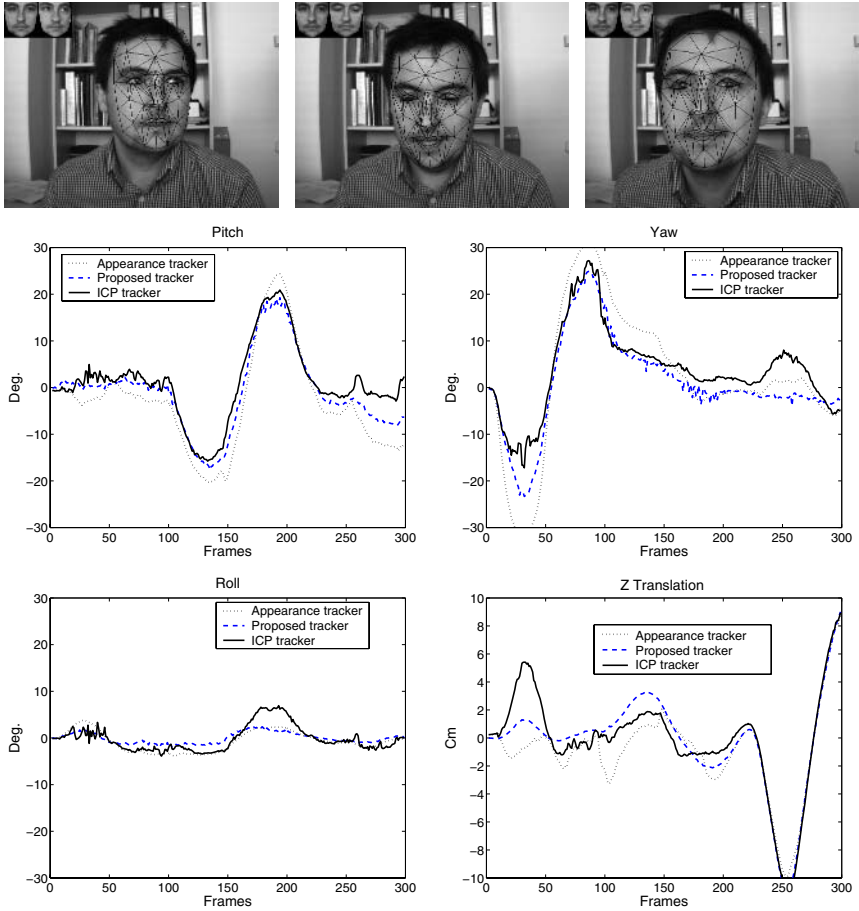


Fig. 6. Top: Tracking the face and facial feature in a 300-frame long sequence using the proposed tracker. Only frames 22, 179, and 255 are shown. **Bottom:** The relative 3D face motion obtained with the three trackers. The dotted curves correspond to the appearance-based tracker, the dashed ones to the proposed framework, and the solid ones to the ICP algorithm.

Figure 6 (Bottom) displays the computed relative 3D face motions obtained with the three methods. This figure displays the three angles and the in-depth translation. The dotted curves correspond to the appearance-based tracker, the dashed ones to the proposed framework, and the solid ones to the ICP algorithm.

From these curves, we can see that the proposed framework has outperformed the appearance-based tracker since the curves become close to those computed by the ICP algorithm - the ground-truth data. In this case, the first facial surface used by the ICP algorithm contained about 20000 3D points. Figure 7 displays the computed relative 3D face motions (pitch and yaw angles) obtained with another video sequence.

Since the ICP algorithm works with rigid surfaces the faces depicted in the sequences of Figures 6 and 7 were somehow neutral. Figure 8 displays the computed relative 3D

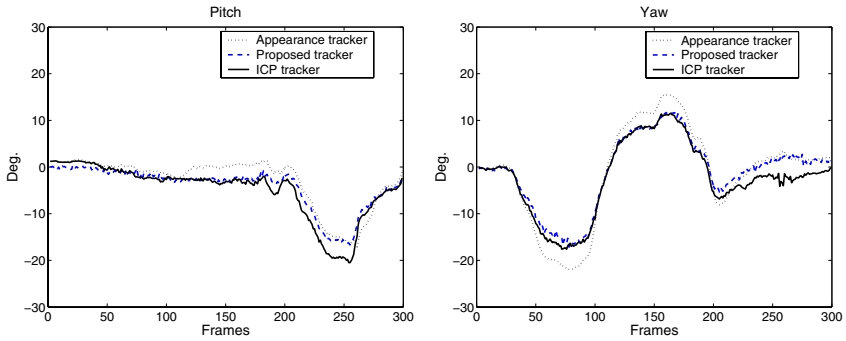


Fig. 7. The relative 3D face motion associated with another video sequence

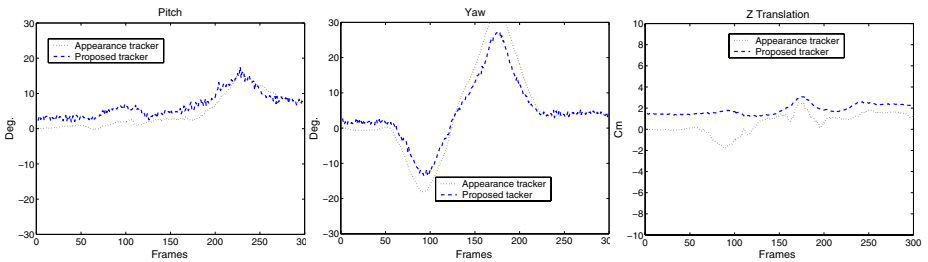


Fig. 8. The relative 3D face motion associated with a video sequence depicting simultaneous head motions and facial expressions

face motions (the three out-of-plane parameters) obtained with another video sequence depicting simultaneous head motions and facial expressions. For this sequence we have not used the ICP algorithm since the facial surface undergoes a rigid and large non-rigid motion. As can be seen, the motion parameters have been improved by using online stereo data.

7 Conclusion

In this paper, we have proposed a robust 3D face tracker that combines the advantages of both appearance-based trackers and 3D data-based trackers while keeping the CPU time very close to that required by real-time trackers. Experiments on real video sequences indicate that the estimates of the out-of-plane motions of the head can be considerably improved by combining a robust 3D-to-3D registration with the appearance model. Although the joint use of 3D facial data and the ICP algorithm as a 3D head tracker could be attractive, the significant computational cost of the ICP algorithm prohibits real-time performance.

References

1. Moreno, F., Tarrida, A., Andrade-Cetto, J., Sanfeliu, A.: 3D real-time tracking fusing color histograms and stereovision. In: IEEE International Conference on Pattern Recognition. (2002)
2. Cascia, M., Sclaroff, S., Athitsos, V.: Fast, reliable head tracking under varying illumination: An approach based on registration of texture-mapped 3D models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **22** (2000) 322–336
3. Ahlberg, J.: An active model for facial feature tracking. *EURASIP Journal on Applied Signal Processing* **2002** (2002) 566–571
4. Matthews, I., Baker, S.: Active appearance models revisited. *International Journal of Computer Vision* **60** (2004) 135–164
5. Malassiotis, S., Srinivasan, M.G.: Robust real-time 3D head pose estimation from range data. *Pattern Recognition* **38** (2005) 1153–1165
6. Besl, P., McKay, N.: A method for registration of 3-D shapes. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **14** (1992) 239–256
7. Ahlberg, J.: CANDIDE-3 - an updated parametrized face. Technical Report LiTH-ISY-R-2326, Department of Electrical Engineering, Linköping University, Sweden (2001)
8. Xiao, J., Baker, S., Matthews, I., Kanade, T.: Real-time combined 2D+3D active appearance models. In: IEEE Int. Conference on Computer Vision and Pattern Recognition. (2004)
9. Jepson, A., Fleet, D., El-Maraghi, T.: Robust online appearance models for visual tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **25** (2003) 1296–1311
10. Dornaika, F., Davoine, F.: On appearance based face and facial action tracking. *IEEE Transactions on Circuits and Systems for Video Technology* (In press)
11. Cootes, T., Edwards, G., Taylor, C.: Active appearance models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **23** (2001) 681–684
12. Dornaika, F., Sappa, A.: Appearance-based tracker: An evaluation study. In: IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance. (2005)
13. Chetverikov, D., Stepanov, D., Kresk, P.: Robust Euclidean alignment of 3D point sets: the trimmed iterative closet point algorithm. *Image and Vision Computing* **23** (2005) 299–309
14. Fitzgibbon, A.: Robust registration of 2D and 3D point sets. *Image and Vision Computing* **21** (2003) 1145–1153
15. Fischler, M., Bolles, R.: Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communication ACM* **24** (1981) 381–395
16. Rousseeuw, P., Leroy, A.: *Robust Regression and Outlier Detection*. John Wiley & Sons, New York (1987)
17. Horn, B.: Closed-form solution of absolute orientation using unit quaternions. *J. Opt. Soc. Amer. A* **4** (1987) 629–642

Curve Mapping Based Illumination Adjustment for Face Detection

Xiaoyue Jiang¹, Tuo Zhao², and Rongchun Zhao¹

¹ College of Computer Science, Northwestern Polytechnical University,
Xi'an, China, 710072

xiaoyuejiang@mail.nwpu.edu.cn

² School of Mechatronic Engineering, Northwestern Polytechnical University,
Xi'an, China, 710072

Abstract. For the robust face detection, illumination is considered as one of the great challenges. Motivated with the adaptation of the human vision system, we propose the curve mapping (CM) function to adjust the illumination conditions of the images. The lighting parameter of CM function is determined by the intensity distribution of the images. Therefore the CM function can adjust the images according to their own illumination conditions adaptively. The CM method will abandon no information of the original images and bring no noises to the images. But it will enhance the details of the images and adjust the images to the proper brightness. Consequently the CM method will make the images more discriminative. Experimental results show that it can improve the performance of the face detection with the CM method as a lighting-filter.

1 Introduction

Face detection has been well regarded as challenging problem in the vision community. Due to variations caused by pose, expression, occlusion, illumination or lighting the distribution of face subject is highly nonlinear, and thus makes the detection tasks extremely difficult. Among these variations, illumination and pose changes are regarded as most critical factors for robust face detection. Recently, view-based framework has been widely used to reduce the variances caused by pose changes [1,2,3].

The methods to deal with the illumination variance can be divided into three types. The first type is to enhance the image. This type of method is focused in adjusting the images to reduce the influence of illumination variance, e.g. the histogram equalization (HE) method. In Rowley's face detection system [1], he proposed a linear model as well as the HE method to do de-lighting work for the images. The second type is to extract the illumination invariant features from images. Through a series of face recognition experiments, Adini [4] proved the high-frequency features which are traditionally considered to be illumination invariant are not enough to represent images across illumination. The quotient image (QI) based methods [5,6] extract the albedo information from images assuming the low-frequency information of the images is a good approximation of

the illumination conditions. However, the assumption is not valid. The reflected light from a subject always contains some abrupt changes (e.g. the edge of shadows or highlight parts) due to the shape of the subject or the character of the light source. The third type method is to model the variance of the images under different illumination conditions. Bellhumeur and Kriegman [7] prove the set of images of a convex subject with a Lambertian surface, illuminated by an arbitrary number of point light sources at infinity, forms a convex polyhedral cone called illumination cone. The dimension of the cone is equal to the number of distinct surface normals. The high dimensionality limits the application of illumination cone. Usually the low dimensional approximation for the cone is applied. With a set of images taken with different lighting conditions Georghiadis, et al[8] constructs a 3D illumination subspace. With the spherical harmonic bases to represent the reflection function Ramamoorthi[9] and Barsi[10] construct a 9D illumination subspace independently. The spherical harmonic is the function about the surface normal. And from an image to recover the surface normal of the subject is till an open problem. Therefore it constrains the application of the spherical harmonic based methods.

In the face detection task, the face should be separated from all the non-face subjects. And always there are no training samples for the faces to be detected. Then the de-lighting methods based on training models from sample images are not suitable for detection. Therefore the image enhancement based de-lighting work will be more effective for detection.

For a certain subject with a certain pose, the direction and intensity of the incident light determines the appearance of the images. The light serves as the amplifier for the subject's surface character. Too dim light will make the images dark and lacking details. But too bright light will also make the image losing its details. Therefore the correction for the images taken under unsuitable illumination conditions is to adjust the intensity of the incident light so that the image will be rich of details and in the proper brightness. Human vision system (HVS) is a precise system that can adapt to the huge variance of the environment illumination. Then based on the mechanics of HVS adaptation, a curve mapping function is proposed to adjust the illumination conditions of the images.

In section 2, we will first briefly review the adaptation of HVS and introduce the curve mapping functions derived from it. Then how to decide the lighting parameters in the mapping function is explained. In section 3 we combine different de-lighting methods with the face detector and compare their effect on the face detection. At last the conclusion is drawn in section 4.

2 Curve Mapping Function

In the reflection function, the incident light can be considered as the amplifier for the reflectance character. Then the elimination of the illumination influence can be transferred to a mapping question. If we can find a suitable mapping function that can adjust the intensity of the incident light, then the influence of the illumination will be alleviated greatly. The human visual system (HVS)

is the best and the most complicated vision system. The system can adjust itself according to the current illumination conditions. We propose a mapping function based on the adaptation of HVS to do the de-lighting work for the images, where the parameter of mapping function is decided according to the illumination conditions of the image itself.

2.1 Adaptation of HVS

The accurate adaptation of HVS is achieved through the cooperation of mechanical, photochemical, and neural processing in the vision system. The pupil, the photoreceptor (rod and cone) systems, bleaching and regeneration of receptor photopigments, and the changes in neural processing all play a role in visual adaptation[11]. According to the results from electro-physiology, the photoreceptor take the main charge in the procedure of adaptation and the photoreceptor can be modeled as the function of input intensity[12]

$$V = \frac{I}{\alpha I + \beta} V_{max} \quad (1)$$

where V is the potential produced by cones; parameter $\alpha > 0$ and $\beta > 0$; V_{max} determines the maximum range of the output value. The semi-saturated parameter λ is an important hidden parameter in Eq.1. It decides the value of I at which the output of the function gets the value $V = V_{max}/2$. Then the semi-saturated parameter λ is

$$\lambda = \frac{\beta}{2 - \alpha} \quad (2)$$

For the semi-saturated parameter, it should satisfy $\lambda \geq 0$, then we have $\alpha < 2$. Specially when $\alpha = 1$, the semi-saturated parameter $\lambda = \beta$, i.e. when $I = V$, gets the value of $V_{max}/2$.

2.2 Mapping Function

Applying the photoreceptor model to adjust the illumination conditions of images, we need to add some constraints to the original formula so that the mapping function can meet the requirement to show images.

The input images is in the range of $[0, 255]$, and the output mapped image should also be in the range of $[0, 255]$. Therefore we set the maximum range $V_{max} = 255$. For the pixels with the value of 0 or 255 in the original image, we will keep their value in the corrected image. The reason is that these pixels represent the darkest and brightest points in the image respectively. And the corrected image should show the details in the largest range, i.e. $[0, 255]$. Therefore the transform function $f(I) = I/(\alpha I + \beta) \in [0, 1]$ should satisfy the following equations

$$f(0) = 0 \quad (3)$$

$$f(255) = 1 \quad (4)$$

Then we can get the relationship between α and β , that is

$$\frac{\beta}{1 - \alpha} = 255 \tag{5}$$

The pixels in the original image are in the range of $[0, 255]$, i.e. $I \in [0, 255]$, then we have

$$I - 255 \leq 0 \tag{6}$$

Based on Eq.6, we can deduce

$$\alpha(I - 255) + 255 \leq 255 \tag{7}$$

Then Eq.7 can be rewritten like

$$\alpha I + 255(1 - \alpha) \leq 255 \tag{8}$$

with Eq.5 and Eq.8, we get

$$\alpha I + \beta \leq 255 \tag{9}$$

from Eq.9 and $I \in [0, 255]$, we have

$$0 < \beta \leq 255 \tag{10}$$

$$0 \leq \alpha < 1 \tag{11}$$

we let $\beta = 255/k(1 \leq k)$ then parameter α can be $(k - 1)/k$, therefore we unify these two parameters and we can rewrite the mapping function as following

$$F_d(I) = \frac{I}{\frac{k_d-1}{k_d}I + \frac{255}{k_d}} \times 255 \tag{12}$$

where the lighting parameter $k_d \geq 1$. Fig.1 gives out the curves of mapping function $F_d(I)$ with lighting parameter k_d getting different value. When $k_d = 1$, the mapping function does not change the original images. While $k_d > 1$, the mapping function will enhance the darker part of original images. The mapping function reaches to the inflexion point when I gets the value of $255/(\sqrt{k_d} + 1)$. That is to say when $I \in [0, 255/(\sqrt{k_d} + 1))$, the slope of the mapping function is larger than 1 and when $I \in (255/(\sqrt{k_d} + 1), 255]$, the slope is smaller than 1.

When the illumination is too dim or too bright, the image will be under-exposure or over-exposure. In the under-exposure situation, the image is darker and the details are lost. The character of the subject is suppressed in the image due to the less amount of light. To recover the information in the darker part, a suitable amount of lighting should be added to that part so that the character of the subject can be represented in a suitable scale. The slope of the mapping function $F_d(I)$ is larger than 1 in the darker domain, therefore the relative value between pixels and the absolute value of every pixel in that region will be enlarged. As a result the contrast and brightness of the darker region are both

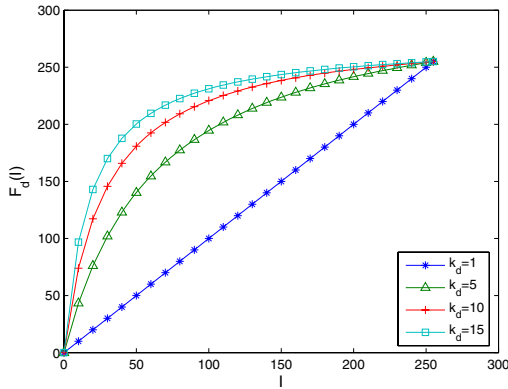


Fig. 1. the curves of mapping function $F_d(I)$

enhanced. The mapping function $F_d(I)$ can take the role to adjust the illumination of under-exposure condition. For the over-exposure situation, the lighting magnifies the character of the subject too large to display. That is the pixels in the brighter region have larger absolute value, but the differences between pixels are suppressed in images. Therefore the correction function should reduce the light amount and at same time enlarge the differences between pixels. Then we can apply the mapping function $F_d(I)$ to map the inverse image, i.e. $(255 - I)$, and get the mapping function in the over-exposure situation,

$$\begin{aligned} F_b(I) &= 255 - F_d(255 - I) \\ &= 255 - \frac{255 - I}{\frac{k_d - 1}{k_d}(255 - I) + \frac{255}{k_d}} \times 255 \end{aligned} \quad (13)$$

where the lighting parameter $k_b > 1$. The mapping function $F_b(I)$ can reduce the brightness of the images and enhance the contrast of the brighter parts.

2.3 Lighting Parameter Decision

In the mapping functions, the parameter k_d and k_b can be decided according to the illumination conditions of the images. Considering of the semi-saturated parameter λ , it decides the point at which the input value of I will be mapped to the middle of the output range, i.e. $V_{max}/2$. In the mapping function $F_d(I)$, the semi-saturated parameter is

$$\lambda_d = \frac{\beta}{2 - \alpha} = \frac{255}{k_d + 1} \quad (14)$$

Similarly the semi-saturated parameter of the mapping function $F_b(I)$ is

$$\lambda_b = \frac{255}{k_b + 1} \quad (15)$$

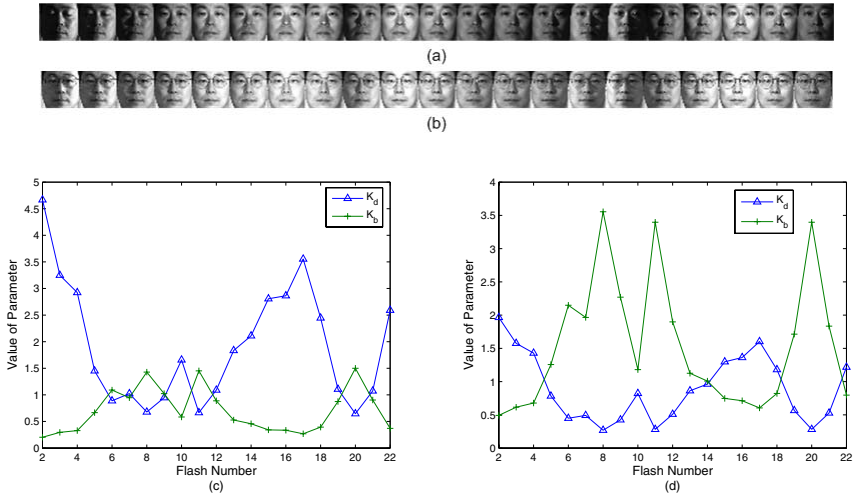


Fig. 2. The value of lighting parameter k_d, k_b for the images under different lighting conditions. (c), (d) are the curves of lighting parameters for the set of images in (a) and (b), respectively.

If an image is rich of detail, it should take use of all the value from 0 to 255 to express all the information. An image whose intensity values concentrate only in a little subset of $[0, 255]$ cannot show out many details of the subject. Therefore we should make the image take as many as possible values to express its details. Then 50% of all the pixels should take at least half of $[0, 255]$ to represent the information. Consequently we can find out an input value $I_h \in [0, 255]$ that satisfies the following equation,

$$\min_{I_h} \left(\sum_{i=0}^{I_h} h(i) \geq \frac{N}{2} \right) \tag{16}$$

where $h(i)$ is the number of pixels in the i th bin of the histogram, N is the total number of the pixels in an image. And if we let I_h equal to the semi-saturated parameter λ_d , the mapping function will map I_h to 128 and stretch the denser part of the image histogram. Then the parameters can be decided as

$$k_d = \frac{255}{I_h} - 1 \tag{17}$$

$$k_b = \frac{255}{I'_h} - 1 \tag{18}$$

where I'_h is decided through counting the histogram of the inverse image with Eq.16. Due to the constraints for k_d and k_b , I_h and I'_h should be smaller than 128. If $I_h > 128$, the image is over-exposure and we get $k_d < 1$ and $k_b > 1$. Therefore we apply the mapping function $F_b(I)$ to adjust the image. Correspondingly if $I'_h > 128$, function $F_d(I)$ is applied. Then the mapping function can be written as

$$F(I) = \begin{cases} F_d(I) & \text{if } k_d > 1 \\ F_b(I) & \text{if } k_b > 1 \\ I & \text{if } k_b = k_d = 1 \end{cases} \quad (19)$$

The parameter k_d and k_b have their physical meaning for the image. They represent the uniformity of the illumination conditions. We choose a set of images from PIE database[14]. From left to right Fig.2(a) are the images taken with no ambient light and the flash No.2 to No.22 firing one by one. From left to right Fig.2(b) are the images taken with ambient light on and flash No.2 to No.22 firing one by one. From the parameter curves, we can see that the more obvious the side illumination is, the larger the parameter value is. With the larger parameter k_d or k_b , the mapping function will have larger slope in the region of $[0, \frac{255}{\sqrt{k_d+1}})$ or $(255\frac{\sqrt{k_b}}{\sqrt{k_b+1}}, 255]$. As a result the correction for the image will be greater. Also we can find that when $k_d > 1$, the image is under-exposure; when $k_b > 1$, the image is over-exposure. Therefore the adjustment for the image can be done according to the image's own illumination conditions.

3 Experiments

In the experiment, we first compare the illumination adjustment results from the view of image appearance. Then we put different de-lighting methods into the face detector and compare the contribution of these de-lighting methods for face detection.

3.1 Illumination Corrected Images

In Fig.3(a), we give out some results of different de-lighting methods which are the histogram equalization (HE), Rowley's linear de-lighting method (RLD), quotient image (QI) method, and the proposed curve mapping (CM) method. In order to compare these results more objectively, we also give out the histogram and edge image for every image. HE method can stretch the denser part of the histogram according to its intensity distribution; however, it leaves much great noise in the image. In RLD method the illumination condition is estimated by a linear model and then the estimated illumination is subtracted from the original image, at last HE is applied to enhance the contrast. The subtraction does not satisfy the reflection theory therefore RLD cannot remove the lighting influences completely. Although QI method can remove the lighting influence, it is at the cost of losing the low frequency information and enhancing some noises. In the result of QI method only leave some edges that contain the original edges of the face as well as those caused by the light. Seen in the CM result, the image is brightened and the contrast is enhanced. Moreover CM does not introduce much great noise like HE or QI method. From the CM results in Fig.3(b), we can see that the CM method adjusts the images based on their own illumination condition. CM can adjust the images to the suitable brightness and enhance their details. Then the corrected images will be more discriminative.

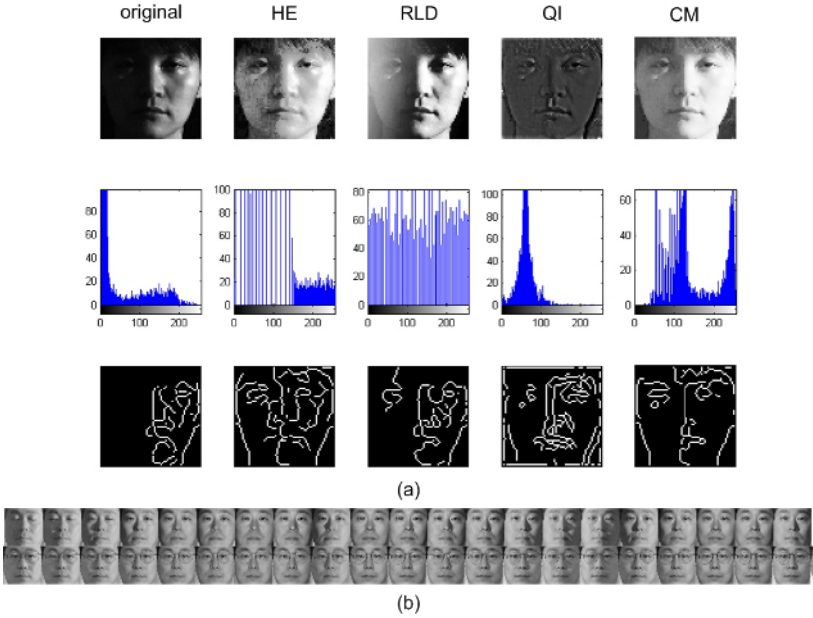


Fig. 3. (a) some results of different illumination correction method. (From up to down are the image, histogram and edge image respectively); (b) CM illumination corrected result of Fig2 (a), (b).

3.2 Detection Results

The face detector is Gabor feature based boosting chain detector. We first extract the Gabor features from images, and then apply the boosting chain method[13] to select the most discriminative features and at the same time construct the classifier based on the selected features. The de-lighting methods are combined with the detector as the pre-filter. That is all the image is processed with the de-lighting method first and then put to the detector.

More than 12000 images without faces and 10000 faces images are collected by cropping from various sources, such as AR[15], Rockfeller, FERET[16], BioID and from WEB. Most face in the training set have the variation of pose and lighting. A total number of about 80000 face samples with size of 32×32 are generated from the 10000 face images by following random transformation: mirror, four-direction shift with 1 pixels, in-plane rotation within 15 degrees and scaling within 20% variations. We randomly select 20000 face samples and 20000 non-face for training. Altogether we train out 5 detectors with different de-lighting methods as pre-filter, they are the raw (no de-lighting), HE, RLD, QI and CM.

We first probe these detectors on PIE database[14]. We separate the PIE database into 4 subsets according to the illumination conditions of the images. Subset1 are the images taken under extreme lighting condition. Subset2 are images taken under the ambient light. Subset3 are images taken with only different flashes. Subset4 are images taken with different flashes and the ambient light.



Fig. 4. The samples of PIE subsets

Table 1. Illumination condition of every PIE subset

	Ambient Light	Flash No.
Subset1	Yes	2,3
	No	2~4 and 19~22
Subset2	Yes	None
Subset3	No	5~18
Subset4	Yes	4~22

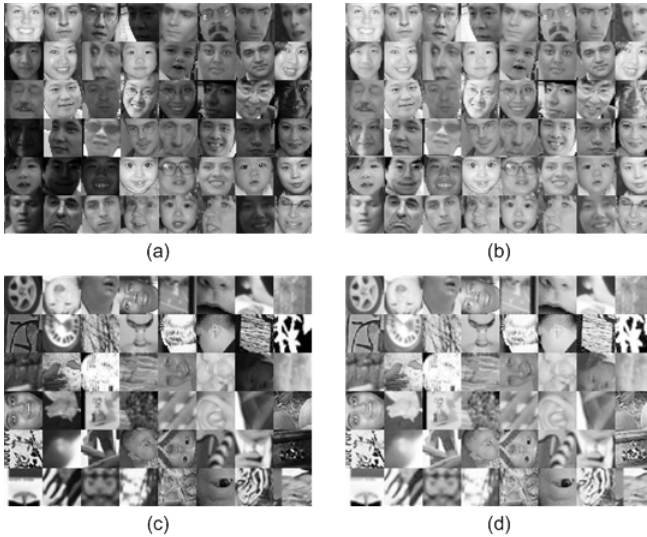
The sample images of every subset are given in Fig.4. The details of every subset are listed in Table1. The detection results are given in Table 2.

From the detection result, we can see that with the influence of illumination become greater the detection rate is reduced. The illumination for the images in subset 2 is the ambient light, and the detection result on subset 2 is better than that on other subset. On images of subset 1 which are influenced by illumination most badly, the detection rate is lowest whatever the lighting filter is. HE, RLD and QI methods can alleviate the influence of illumination in certain degree. However, HE method introduces some great noise; RLD cannot remove the illumination influence completely and at the same time brings much great noise due to the limitation of HE. QI adjusts the illumination at the cost of abandoning much low frequency information. Consequently the detection rates of the detectors with these pre-filters have no great improvement. CM method can adjust the images according to the lighting condition of their own. It can enhance the contrast and enrich the details without introducing much noise. Therefore an image taken under extreme illumination conditions can be recovered to a normal illumination condition by CM method. And in the detection result the improvement is shown.

From the detection result, we can see that with the influence of illumination become greater the detection rate is reduced. The illumination for the images in subset 2 is the ambient light, and the detection result on subset 2 is better than that on other subset. On images of subset 1 which are influenced by illumination most badly, the detection rate is lowest whatever the lighting filter is. HE, RLD and QI methods can alleviate the influence of illumination in certain degree. However, HE method introduces some great noise; RLD cannot remove the illumination influence completely and at the same time brings much great noise due to the limitation of HE. QI adjusts the illumination at the cost of abandoning much low frequency information. Consequently the detec-

Table 2. Detection rate on the PIE datasets

	Subset1	Subset2	Subset3	Subset4
Raw	535/63	436/2	887/28	1260/32
HE	563/35	437/1	898/17	1274/18
RLD	537/61	436/2	891/24	1262/30
QI	575/23	433/5	902/13	1273/19
CM	597/2	438/0	914/1	1291/1

**Fig. 5.** (a), (c) some face and non-face samples of probe set; (b), (d) the CM result of (a), (c)

tion rates of the detectors with these pre-filters have no great improvement. CM method can adjust the images according to the lighting conditions of their own. It can enhance the contrast and enrich the details without introducing much noise. Therefore an image taken under extreme illumination conditions can be recovered to a normal illumination condition by CM method. And in the detection result the improvement is shown.

Then we test these detectors on more general images which are taken in more natural situations. We randomly select 9024 positive samples and 9315 negative samples from the 80000 face images and 12000 non-face images that have just introduced In Fig.5 we give some face and non-face samples of the probe set and their results of CM adjustment. Seen from the results, the CM method is independent from the content of the images. It adjusts the images only according to the illumination of the images. We give out the ROC curves of the detectors with different pre-filter in Fig.6. We can see that the performance of the detector with the CM method as the pre-filter is better than the others. It is because the CM method will not abandon some information of the original image or bring

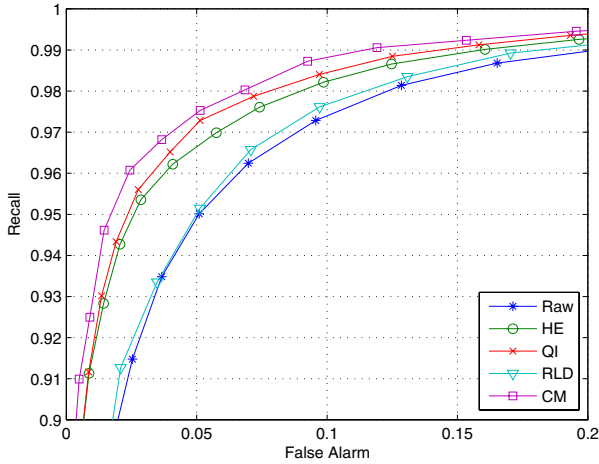


Fig. 6. The ROC curves of detectors with different de-lighting methods as pre-filter

some noises to the images, but enhance some information in the image. As a result, the CM method will adjust the images to be more discriminative.

4 Conclusion

There are three different approaches to deal with the illumination variance. For the detection task, the image enhancement based method is the most suitable one. It is because there are no training images for what will be detected. According to the reflection theory, the incident light can serve as the amplifier for the character of the subject. Then adjusting the images is turn to adjust the incident light. Motivated with the adaptation of HVS, we proposed a curve mapping function. This mapping function can adjust the images to a more normal illumination condition and enrich their details which are suppressed in the original images. The adjustment is done according to the illumination situation of the image its own. With suitable brightness and abundant details, the adjusted images are more discriminative. In the experiments, the performance of CM filter is much better than the others.

The illumination variance is only one of the factors that make the robust face detection difficult. To achieve more robust face detection, we need to consider of the pose, expression variance and so on.

References

1. Rowley, H. A., Baluja, S., and Kanade, T: Neural network-based face detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, no. 1, (1998) 22-38
2. Schneiderman, H. and Kanade, T: A Statistical Method for 3D Object Detection Applied to Faces and Cars. *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, (2000)

3. Li, S. Z., Zhu, L., Zhang, Z. Q., et al: Statistical Learning of Multi-View Face Detection. Proc. of the 7th European Conf. on Computer Vision. (2002)
4. Adini, Y., Moses, Y. and Ullman, S.: Face recognition: the problem of compensating for changes in illumination direction. IEEE Tran. Pattern Recognition and Machine Intelligence, vol.19, no.7, (1997)721-732
5. Shashua, A., and Riklin-Raviv, T.: The Quotient Images: Class-based Re-Rendering and Recognition with Varying Illuminations. IEEE Tran. Pattern Recognition and Machine Intelligence, vol.23, no.2, (2001)129-139
6. Wang, H., Li, S. Z., and Wang, Y.: Generalized quotient image. IEEE Conference on Computer Vision and Pattern Recognition, (2004)
7. Belhumeur, P. and Kriegman, D., "What is the set of images of an object under all possible lighting conditions?" IEEE Conf. Computer Vision and Pattern Recognition, pp.270-277,1996
8. Georghiads, A., Belhumeur, P. and Kriegman, D.: From few to many: illumination cone models for face recognition under variable lighting and pose. IEEE Tran. Pattern Recognition and Machine Intelligence, vol.23, no.6, (2001)643-660
9. Ramamoorthi, R. and Hanrahan, P.: A signal-processing framework for inverse rendering. SIGGRAPH, (2001)117-128.
10. Basri, R. and Jacobs, D. : Lambertian reflectance and linear subspace. IEEE Tran. Pattern Analysis and Machine Intelligence, vol. 25, no. 2, (2003)218-233
11. Ferwerda, J. A., Pattanaik, S. N., Shirley, P., and Greenberg, D. P.:A model of visual adaptation for realistic image synthesis. Proceedings of the 23rd annual conference on Computer graphics and interactive techniques (1998)
12. Reinhard, E., Stark, M., Shirley, P., and Ferwerda, J.: Photographic tone reproduction for digital images. ACM Transactions on Graphics, vol. 21, no. 3, (2002)267-276
13. Xiao, R., Li, M., Zhang, H.: Robust multipose face detection in Images. IEEE trans. on Circuits and Systems for video technology, vol. 12 no.1, (2004) 31-41
14. Sim, T., Baker, S., and Bsat, M.: The CMU Pose, Illumination, and expression (PIE) database. Processing of the IEEE International Conference on Automatic Face and Gesture Recognition, (2002)
15. Martinez, A.M. and Benavente, R: The AR Face Database. CVC Technical Report #24, (1998)
16. Phillips, P.J., Wechsler, H., Huang, J., and Rauss, P.J.: The FERET database and evaluation procedure for face-recognition algorithms. Image and Vision Computing, 16(5), (1998) 295-306

Common Image Method(Null Space + 2DPCAs) for Face Recognition

Hae Jong Seo, Young Kyung Park, and Joong Kyu Kim

School of Information and Communication Engineering, SKKU
300, Cheon-Cheon, Jang-An, Suwon, Kyung-Ki, Korea 440-746
rokaf539@skku.edu

Abstract. In this paper, we present a new scheme called Common Image method for face recognition. Our method has a couple of advantages over the conventional face recognition algorithms; one is that it can deal with the Small Sample Size(SSS) problem in LDA, and the other one is that it can achieve a better performance than traditional PCA by seeking the optimal projection vectors from image covariance matrix in a recognition task. As opposed to traditional PCA-based methods and LDA-based methods which employ Euclidean distance, Common Image methods adopted Assemble Matrix Distance(AMD) and IMage Euclidean Distance(IMED), by which the overall recognition rate could be improved. To test the recognition performance, a series of experiments were performed on CMU PIE, YaleB, and FERET face databases. The test results with these databases show that our Common Image method performs better than Discriminative Common Vector and 2DPCA-based methods.

1 Introduction

Principle Component Analysis(PCA) and other PCA-based techniques are well-known schemes for image representation and recognition. Recently, Eigenface for recognition [2] based on PCA was also proposed by Turk and Pentland in 1991, and since then, Eigenface-based face recognition schemes have been extensively investigated. The key idea behind the Eigenface method is to find the optimal directions in the sample space that will maximize the total scatter across all images. More recently, a novel image representation and recognition technique, two-dimensional PCA(2DPCA) [1] has been proposed by Yang et.al.(2004). The 2DPCA is based on 2D image itself without the need to be transformed into a vector, i.e., the so-called image covariance matrix is constructed directly using the 2D image matrices. Hui Kong et.al.(2005) [17] theoretically proved that 2DPCA should always outperform the PCA. Although 2DPCA performs better than conventional PCA, the shortcoming of it is that it only reflects variations between rows of images, not including the columns. In 2005, a novel method called diagonal principal component analysis(DiaPCA), which is capable of taking account of variations of rows as well as those of columns of images, was proposed by Daoqiang Zhang et.al. [18]. While the above-mentioned PCA-based methods are efficient for feature extraction, they have an intrinsic limitation that they do not reflect the within-class scatter of subjects.

On the other hand, there have been other developments for feature extraction and recognition such as Linear/Fisher Discriminant Analysis(LDA/FDA) [3, 4] scheme. This method can overcome the above-mentioned limitation of the PCA-based methods by finding out the best projection space so that the ratio of the between-class scatter to the within-class scatter is maximized. Even though it can solve the problem PCA-based methods have, it also has its own intrinsic limitation that the within-class covariance must be nonsingular when we do not have enough images for training, i.e., when the “Small Sample Size(SSS)problem or singularity problem [12]” arise. In order to solve this problem, many approaches extending LDA have been proposed in recent years. These are PCA+LDA [6], N-LDA [7], PCA+Null Space [3], Direct-LDA [8], Random sample LDA [9], and Dual-space LDA [10]. Among these LDA extensions, N-LDA(NullSpace-LDA), PCA+NullSpace, and Direct LDA incorporate the idea of nullspace of the within-class scatter. Lately, a novel method which is called Discriminative Common Vector(DCV) method [5] was proposed. It is a nullspace-based method which can efficiently extract discriminative common vectors which well represent each subject by projecting the original sample vector directly into the optimal projection space. Since this method incorporates PCA in order to obtain the optimal projection space, it can not avoid the disadvantage that the 2D face image matrices must be transformed into 1D image vectors [1]. This disadvantage, as mentioned above, can be overcome by applying two-dimensional PCAs(2DPCA, DiaPCA, and DiaPCA+2DPCA).

In this paper, a framework of Common Image method, which simultaneously incorporates advantages of the nullspace method and two-dimensional PCAs, is proposed for face recognition. The proposed framework consists of three different types of scheme: Nullspace+2DPCA, Nullspace+DiaPCA, and Nullspace+DiaPCA+2DPCA. By applying these proposed schemes, the existing limitations of both LDA and PCA based methods are simultaneously resolved, thus the performance is expected to improve to a substantial extent. We adopt the Image Euclidean Distance(IMED) [11] and Assembled Matrix Distance(AMD)metric [14] instead of traditional Euclidean distance as a similarity measure. Moreover, since almost all face recognition methods encounter difficulties under varying lighting conditions, the SSR (Single Scale Retinex) [19] as a pre-processing step is additionally employed in order to minimize the effect of illumination.

The rest of this paper is organized as follows. In Section 2, three types of Common Image methods for face recognition are described. Pre-processing step is presented in Section 3. In Section 4, the experimental results are provided using images in the CMU PIE [13], YaleB [15], and FERET [16] databases. Finally, a conclusion is given in Section 5.

2 Common Image Analysis

The method we propose is based on Discriminative Common Vector(DCV) method [5] and two-dimensional PCA(2DPCA) methods [1, 18]. Discriminative common vector is a common feature extracted from each class in the training set

by eliminating the differences of the samples in each class. In order to obtain the common vectors, the within-class scatter matrix of all classes is used instead of given class's own scatter matrix. In our case, after obtaining common vectors of each class, we transform vectors into matrices so that two-dimensional PCAs can be applied afterwards. Two-dimensional PCAs can directly extract optimal feature matrices of each class from images, thus has many advantages over classical PCA. We introduce three approaches for obtaining optimal features according to the variations of two-dimensional PCA after projecting the training set onto the null space of within-scatter matrix.

2.1 NullSpace + 2DPCA(N-2DPCA)

Let a training set consists of C classes, where each class includes N samples, and let A_i be a d -dimensional column vector($d = m(\text{height}) \times n(\text{width})$). Now that this method relies on the NullSpace method, it is required to satisfy that $d > NC$. In this case, the within scatter matrix S_W , the between scatter matrix S_B , and, the total scatter matrix S_T are defined as

$$S_W = \sum_{j=1}^C \sum_{i=1}^N (A_i^j - \mu_j)^T (A_i^j - \mu_j), \quad S_B = \sum_{j=1}^C N(\mu_j - \mu)^T (\mu_j - \mu) \quad (1)$$

$$S_T = \sum_{j=1}^C \sum_{i=1}^N (A_i^j - \mu)^T (A_i^j - \mu) = S_W + S_B \quad (2)$$

where, μ is the mean of all samples, and μ_j is the mean of N samples in the j th class.

In order to obtain common vectors of each class, we project the face samples onto the null space of S_W and use orthogonal complement of the null space of S_W . Let R_d be the sample space, V be the range space of S_W , V^\perp be the null space of S_W . Then, since $R_d = V \oplus V^\perp$, every training image $A_i^j \in R^d$ can be decomposed into a unique form.

$$A_i^j = PA_i^j + \bar{P}A_i^j \quad (3)$$

In (3), the P and \bar{P} are the orthogonal projection operators onto V and V^\perp , respectively.

When any sample from the training set is projected onto the null space of S_W , we can obtain a unique common vector.

$$A_{com}^j = A_i^j - PA_i^j = \bar{P}A_i^j, \quad i = 1, \dots, N., j = 1, \dots, C. \quad (4)$$

Now, we transform the common vectors of all class into $m \times n$ images in order to apply 2DPCA. Based on the transformed common images, the so-called *image covariance matrix* [1] is defined as

$$G_T(I_{com}) = \frac{1}{C} \sum_{j=1}^C (I_{com}^j - \bar{I}_{com})^T (I_{com}^j - \bar{I}_{com}) \quad (5)$$

where I_{com}^j is the common image from A_{com}^j and $\bar{I}_{com} = 1/C \sum_j I_{com}^j$ is the mean common image.

By maximizing the image scatter criterion,

$$J(X) = X^T G_T X \quad (6)$$

we finally obtain the optimal projection space X_{opt} . The optimal projection vectors, $X_{opt} = [X_1, X_2, \dots, X_d]$ is the eigenvectors of G_T corresponding to the d largest eigenvalues. These projection vectors are also subject to the orthonormal constraints. Now, X_{opt} are used for the extraction of the following features.

$$Y_j = I_{com}^j X_{opt}, \quad j = 1, 2, \dots, C. \quad (7)$$

We call the features Y_{j_s}' the discriminative common matrices and they will be used for identification of test images.

To recognize a test image A_{test} , it is first transformed into d -dimensional column vector. Then this column vector is projected onto the null space of S_W of the training set.

$$A_{projected} = A_{test} - P A_{test} = \bar{P} A_{test} \quad (8)$$

The vector projected onto the null space of S_W of the training set is now re-transformed into the $m \times n$ image in order to extract the feature matrix of a test image. The feature matrix of the test image is found by

$$Y_{test} = I_{test} X_{opt} \quad (9)$$

where, I_{test} is the image from $A_{projected}$.

2.2 NullSpace + DiaPCA(N-DiaPCA)

We now apply DiagonalPCA [18] into our Common Image method. Let us consider that we obtained the Common Images from each class by projecting the null space of S_W as in (4). Then, common images are transformed into corresponding diagonal face images. Suppose that there is a $m \times n$ image and usually m (height) is greater than n (width). In this case, we use the method as illustrated in Fig.1 to generate the diagonal image D_{com} for the original common image I_{com} . Then, we define the diagonal image covariance matrix as

$$G_T(D_{com}) = \frac{1}{C} \sum_{j=1}^C (D_{com}^j - \bar{D}_{com})^T (D_{com}^j - \bar{D}_{com}). \quad (10)$$

where D_{com}^j is the diagonal image from I_{com}^j and $\bar{D}_{com} = 1/C \sum_j D_{com}^j$ is the mean diagonal image.

Using (6) and (7), discriminative common matrices Y_{j_s}' are obtained. This method can overcome the intrinsic problem of 2DPCA, which is a unilateral-projection-based scheme, by reflecting information between rows and those between columns.

Given a test image, the rest of the recognition procedure is same as in section 2.1, which is described in (8) and (9).

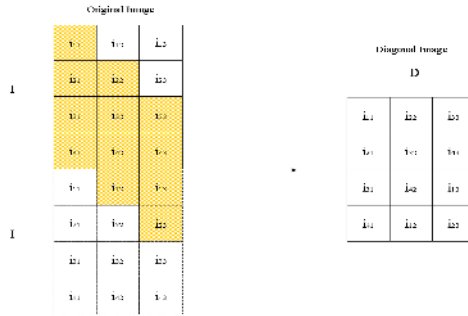


Fig. 1. Illustration of the ways for deriving the diagonal face images [18]

2.3 NullSpace + DiaPCA + 2DPCA(N-DiaPCA+2DPCA)

In this section, DiagonalPCA and 2DPCA are applied in a row. As mentioned in Section 2.2 that m (height) is greater than n (width), diagonal image covariance matrix is first calculated according to the method described in Fig.1 and (11). After obtaining the discriminative common matrix X_{opt} from diagonal PCA, another projection matrix $W_{opt} = [W_1, \dots, W_q]$ of 2DPCA is computed from the q eigenvectors corresponding to the q biggest eigenvalues of the alternative covariance matrix

$$G_T(I_{com}) = 1/C \sum_{j=1}^C (I_{com}^j - \bar{I}_{com})(I_{com}^j - \bar{I}_{com})^T. \tag{11}$$

We first project $m \times n$ common images I_{com}^j onto the $n \times d$ projection space X_{opt} , and then $m \times q$ projection space W_{opt} is used in a next step to finally acquire the $q \times d$ feature matrices,

$$Y^j = W_{opt}^T I_{com}^j X_{opt}, \quad j = 1, 2, \dots, C. \tag{12}$$

Given a test image, by projecting it onto two feature spaces from 2DPCA and DiaPCA in a row, the feature matrix of a test image is given by

$$Y^{test} = W_{opt}^T I_{test}^j X_{opt} \tag{13}$$

2.4 Classification Method

For each of the Common Image method described in Section 2.1, 2.2, and 2.3, which is also summarized in Table 1, the feature matrix of test image Y_{test} is then compared with the discriminative common matrix Y_j of each class. As mentioned in Section 1, unlike the conventional face recognition scheme, we adopted the Assembled Matrix Distance(AMD) [14] metric as a similarity measure. Here, the distance between two feature matrices, $Y_j = (w_{xy}^j)_{m \times d}$ and $Y_{test} = (w_{xy}^{test})_{m \times d}$ is

$$d_{AMD}(Y_j, Y_{test}) = \left(\sum_{y=1}^d \left(\sum_{x=1}^m (w_{yx}^j - w_{yx}^{test})^2 \right)^{1/2p} \right)^{1/p} (p > 0) \tag{14}$$

Table 1. Common Image methods(NullSpace+2DPCAs) for face recognition

Algorithm : Discriminative Common Image Algorithm		
Input: All Training Image I Preprocessing: Standardizing Transform $I \rightarrow I'$ Output: Discriminative Common Image $Y_{opt} \xrightarrow{L2DPCA}, Y_{opt} \xrightarrow{D2DPCA}, Y_{opt} \xrightarrow{L2DPCA+L2DPCA}$		
1. Transform images into vectors $I' \rightarrow A$ 2. Compute the mean of each class μ_j , Construct within scatter matrix $S_w = \sum_{j=1}^C \sum_{i=1}^{M_j} (A_i^j - \mu_j)(A_i^j - \mu_j)^T$ 3. Construct the range space PA_i^j and the null space $\bar{P}A_i^j$ of S_w by eigen value analysis 4. Obtain unique common vector A_{com}^j of each class by projecting the null space of S_w 5. Retransform common vectors A_{com}^j into images I'_{com}		
2D PCA	Dia PCA	Dia PCA + 2D PCA
6. Construct Image Covariance Matrix $G_x(I'_{com})$ 7. Compute the eigenvectors X_i of $G_x(I'_{com})$ corresponding to the d largest eigenvalues 8. Construct the optimal projection space $X_{opt} = [X_1, X_2, \dots, X_d]$ 9. Obtain the Discriminative Common Image $Y_{opt} \xrightarrow{2DPCA} = I'_{com} X_{opt}$	6. Make diagonal images D'_{com} from Images I'_{com} 7. Construct Image Covariance Matrix $G_x(D'_{com})$ 8. Compute the eigenvectors X_i of $G_x(D'_{com})$ corresponding to the d largest eigenvalues 9. Construct the optimal projection space $X_{opt} = [X_1, X_2, \dots, X_d]$ 10. Obtain the Discriminative Common Image $Y_{opt} \xrightarrow{D2DPCA} = I'_{com} X_{opt}$	6. Make diagonal images D'_{com} from Images I'_{com} 7. Construct Image Covariance Matrix $G_x(D'_{com})$ and $G_x(I'_{com})$ 8. Compute the eigenvectors X_i , W_i of $G_x(D'_{com})$ and $G_x(I'_{com})$ corresponding to the d and q largest eigenvalues respectively 9. Construct the optimal projection space $X_{opt} = [X_1, X_2, \dots, X_d]$ and $W_{opt} = [W_1, W_2, \dots, W_q]$ 10. Obtain the Discriminative Common Image $Y_{opt} \xrightarrow{L2DPCA+2DPCA} = W_{opt}^T I'_{com} X_{opt}$

3 Pre-processing

In order to make images suitable for recognition tasks, images are needed to be made as smooth and noiseless as possible. For this purpose, we adopted the Standardizing Transform as a pre-processing method. Suppose there are two $m \times n$ images, then two images can be represented as $A = (a^1, a^2, \dots, a^{mn})$ and $B = (b^1, b^2, \dots, b^{mn})$, and the Image Euclidean distance between two images A and B is expressed as

$$D_{IME}^2(A, B) = \sum_{i,j=1}^{mn} g_{ij}(a^i - b^j)(a^i - b^j) = (A - B)^T G(A - B) \quad (15)$$

where, the metric matrix $G = (g_{ij})_{mn \times mn}$ can efficiently be embedded into two images applying Standardizing transform(ST) [11] directly as:

$$(A - B)^T G(A - B) = (A - B)^T G^{\frac{1}{2}} G^{\frac{1}{2}} (A - B) = (U - V)^T (U - V) \quad (16)$$

where, $U = G^{\frac{1}{2}} A$ and $V = G^{\frac{1}{2}} B$

The ST-processed images U or V are described as Fig.2.(b). Once transformed into smooth and noiseless images, we further get rid of the illumination effects by applying SSR [19] as in (17).

$$I(x, y) = (\log U(x, y) - \log[F(x, y) * U(x, y)]) \quad (17)$$

In(17), “ $*$ ” denotes the convolution operator, $F(x, y)$ is the gaussian smoothing function, and $U(x, y)$ is the ST-processed image. By removing the smoothed im-

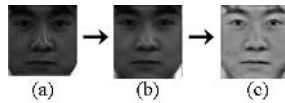


Fig. 2. (a) Original image, (b) the image smoothed by the ST(IMED), (c) the SSR-processed image after ST [19]

age by $F(x, y)$ from the ST-processed image, we finally obtain the SSR-processed image as shown in Fig.2.(c).

4 Experimental Results

The FERET [16], CMU PIE [13], and YaleB [15] databases were used to test our proposed methods.

4.1 Experiments with the FERET Database

The subset that we collected from the FERET face database comprises 612 gray-level frontal view face images from 153 persons. Each person has 4 images (\mathbf{fa} , \mathbf{fb} , \mathbf{fa}_1 , \mathbf{fb}_1) with different facial expressions in different sessions. We pre-processed these images by cropping and normalizing to a size of 92×84 , and aligned the images based on the positions of the eyes. The examples of pre-processed images from FERET databases are shown in Fig.3. In addition to our proposed methods, we also tested the performance of DCV, 2DPCA, DiaPCA, and DiaPCA+2DPCA methods for comparison. The recognition rates were computed by the k -fold strategy [20]: i.e., $k(2 \leq k \leq 3)$ images of each subject are selected for training and the remaining $(4 - k)$ images of each subject are selected for test. The nearest-neighbor algorithm was employed using AMD(Assemble Matrix Distance) for our proposed methods and other two-dimensional PCAs, whereas Euclidean distance has been used for DCV. This process was repeated $(4 - k + 1)$ times and the final recognition rate was computed by averaging the recognition rates from each run. The results are summarized in Table 2. To demonstrate the effect of IMED and AMD, we included the test results, with and without these procedures.

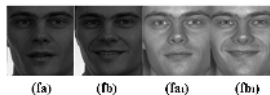


Fig. 3. Sample images for one subject on FERET database

4.2 Experiments with the CMU PIE and YaleB Databases

The CMU PIE database contains 41,368 images obtained from 68 individuals. The database contains image variations according to the illumination, pose, and

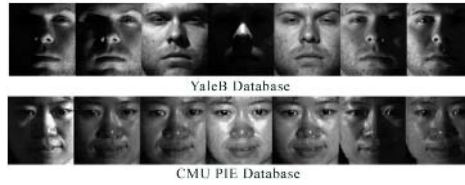


Fig. 4. Sample images for one subject on YaleB and CMU PIE databases

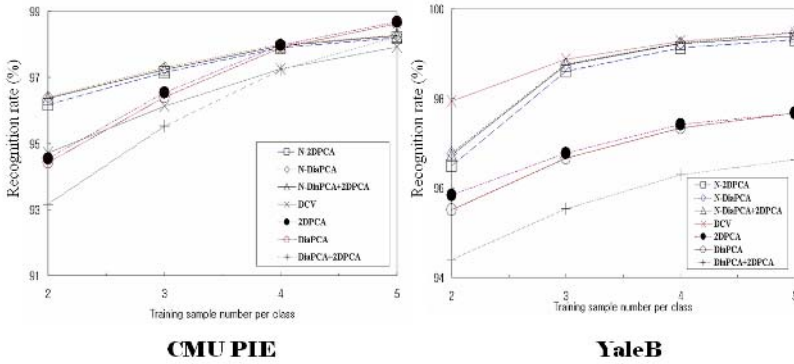


Fig. 5. Performance of proposed methods compared with the state-of-art methods on CMU PIE and YaleB database

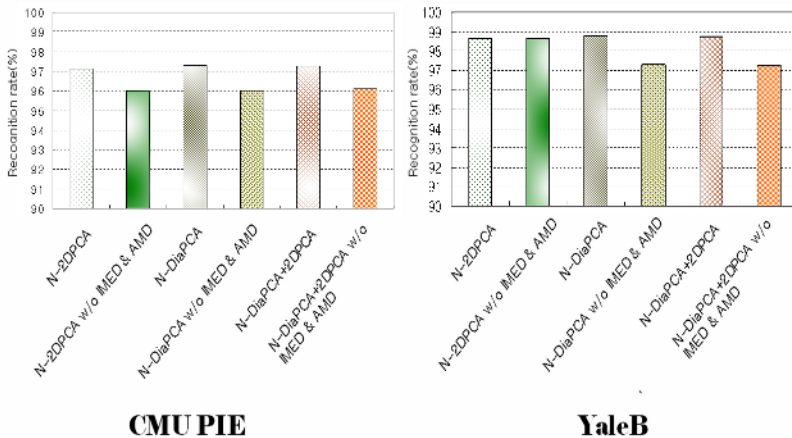


Fig. 6. Performance of proposed methods according to whether they employed AMD and IMED or not on CMU PIE and YaleB database ($N=2$)

facial expressions. We took images of frontal faces with 21 different illumination conditions. Among 68 subjects, we removed one subject because it was not a frontal image. Thus, the total number of images we used for our test is 1,407.

YaleB database contains 5,760 images taken from 10 individuals under 576 viewing conditions(9 poses \times 64 illumination conditions). Here, we used 640 images for 10 subjects representing illumination conditions under the same frontal pose.

We followed the same procedure previously described in the FERET database experiment. Some examples of preprocessed images from CMU PIE and YaleB databases are shown in Fig.4.

In this experiment, the same process as in the experiment in the FERET database was applied except that $k(2 \leq k \leq 5)$ images are selected for training and the remaining($N - k$) images of each subject (CMU PIE: $N=21$, YaleB: $N=6$) are selected for test. The experimental results are depicted in Fig.5 and Fig.6.

4.3 Comparison and Discussion

The dimension of the feature vector in DCV for classification is all reserved to at most $C-1$ (FERET($C=153$), CMU PIE($C=67$), and YaleB($C=10$), where C represents the number of classes as mentioned in Section 2.1). For 2DPCA, DiaPCA, and DiaPCA+2DPCA, the dimensions of reserved feature matrices are 94×15 , 94×16 , and 15×16 respectively. The dimensions of the feature matrices in our proposed methods in this paper are tantamount to those of 2DPCA, DiaPCA, and DiaPCA+2DPCA.

From Table 2 and Fig.5, it is clear that the proposed methods outperform DCV and two-dimensional PCAs, especially when there are few training samples for

Table 2. Comparison of the state-of-art methods on FERET database

Method	N=2		N=3	
	Accuracy(%)	Dimension	Accuracy(%)	Dimension
DCV[5]	37.03	152	46.73	152
2DPCA[2]	38.34	92×15	47.38	92×15
DiaPCA[18]	37.03	92×16	46.73	92×16
DiaPCA+2DPCA[18]	41.50	15×16	50.32	15×16
N-2DPCA	79.19	92×15	86.60	92×15
N-DiaPCA	79.19	92×16	86.27	92×16
N-DiaPCA+2DPCA	79.30	15×16	85.95	15×16

Table 3. Recognition rates on FERET database with and without the IMED and AMD processing

Method	N= 2			N=3	
	Subset 1	Subset 2	Subset 3	Subset 1	Subset 2
N-2DPCA	70.92	86.27	80.39	81.70	91.50
N-2DPCA(w/o IMED and AMD)	53.27	65.36	59.48	76.47	81.70
N-DiaPCA	71.57	85.95	80.07	82.35	90.20
N-DiaPCA(w/o IMED and AMD)	44.12	53.92	50.65	57.52	71.24
N-DiaPCA+2DPCA	71.57	85.95	80.39	81.70	90.20
N-DiaPCA+2DPCA(w/o IMED and AMD)	43.79	54.48	50.33	57.52	71.24

each subject. The underlying reason is that our proposed methods, by efficiently combining the nullspace and two-dimensional PCAs, can extract the optimal feature matrices. Notice that the proposed methods consistently keep higher recognition rates than DCV and two-dimensional PCAs on all of the YaleB, FERET, and CMU PIE databases. One exception is observed when we apply the YaleB database with $N=2$, in which case the recognition rate of DCV is better than our proposed methods. This seems to be due to the fact that the nullspace size of the S_W for YaleB($C=10$) database is comparatively larger than those of the other two FERET($C=153$) and CMU PIE($C=67$) databases. It is known that DCV performs better when the number of classes is smaller, or equivalently when the nullspace size of S_W is larger. For the rest of cases, our proposed methods perform better than DCV, since we take full advantage of two-dimensional PCAs making our algorithm be less sensitive to the size of the nullspace of S_W . On the other hand, the recognition accuracy could not be improved any further, which is an opposing result to Daoquiang Zhang et.al. [18]. Nevertheless, it is clear that N-DiaPCA+2DPCA is more efficient than N-2DPCA and N-DiaPCA when computing eigenvectors because it has a smaller dimension than others. And, this might be an important issue when we consider real-time applications.

Finally, Table 3 and Fig.6 show that the recognition performances of our Common Image methods consistently yield better performance when we applied the AMD and IMED processings, regardless of the databases. This accords with the experimental results of Liwei Wang et.al. [11] and Wangmeng Zuo.et.al [14]. There is, thus, no doubt that the AMD and IMED play significant roles in improving the average recognition rate.

5 Conclusion

In this paper, a framework of Common Image method is proposed for face recognition. The essential idea of the proposed method is to incorporate the idea of the nullspace and two-dimensional PCAs; thus, it not only deals with SSS problem in LDA, but also performs better than the traditional PCA. Experimental results on CMU PIE, YaleB, and FERET databases verify that our proposed methods(N-2DPCA, N-DiaPCA, and N-DiaPCA+2DPCA) are much more accurate than the state-of-art methods(DCV, 2DPCA, DiaPCA, and DiaPCA+2DPCA), especially when there are only small amounts of training images available. We are currently working on simultaneous face identification/authentication problems by a way of applying our Common Image method.

References

- [1] Yang,J., Zhang,D., Frangi,A.F., Yang,J.: Two-Dimensional PCA:A New Approach to Appearance-Based Face Representation and Recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence(2004).
- [2] Turk,M., Pentland,A.: Eigenfaces for Recognition.Journal of Cognitive Neuroscience(1991)

- [3] Belhumeur,P.N., Hespanha,J., Kriegman,D.J.: Eigenfaces vs. Fisherfaces:Recognition Using Class Specific Linear Projection. IEEE Transactionsn Pattern Analysis and Machine Intelligence(1997)
- [4] Zhao,W.: Discriminant Component Analysis for Face Recognition.Int. Conf. on Pattern Recognition(2000)
- [5] Cevikalp,H., Neamtu,M., Wilkes,M., Barkana,A.: Discriminative Common Vectors for Face. IEEE Transactionsn Pattern Analysis and Machine Intelligence(2005)
- [6] Swets,D., Weng,J.: Using discriminant eigenfeatures for image retrieval. IEEE Transactions Pattern Analysis and Machine Intelligence,Vol. 18.(1996)831–836
- [7] Chen,L., Liao,H., Ko,M., Lin,J., Yu,G.:A new lda-based face recognition system which can solve the small sample size problem. Pattern Recognition(2000)
- [8] Yu.H, Yang,J.: A direct lda algorithm for high-dimensional data with application to face recognition.Pattern Recognition,Vol. 34(2001)2067 2070
- [9] Wang,X., Tang,X.: Random sampling lda for face recognition.IEEE International Conference on Computer Vision and Pattern Recognition(2004)
- [10] Wang,X., Tang,X.: Dual-space linear discriminat analysis for face recognition.IEEE International Conference on Computer Vision and Pattern Recognition(2004)
- [11] Wang,L., Zhang,Y., Feng,J.: On the Euclidean Distance of Images.IEEE Transactions on Pattern Analysis and Machine Intelligence,Vol. 26(2004)4 13
- [12] Huang,R., Liu,Q.S., Lu,H.Q., Ma,S.D.: Solving the Small Sample Size Problem of LDA.Proceeding IEEE ICPR(2002)
- [13] Sim,T., Baker,S., Bsat,M.: The CMU Pose,Illumination,and Expression Database.IEEE Transactions on Pattern Analysis and Machine Intelligence,Vol. 25(2003)1615 1618
- [14] Zuo,W., Wang,K., Zhang,D.: Bi-Directional PCA with Assembled Matrix Distance Metric : IEEE ICIP 2005. Vol. 2(2005)958 - 961
- [15] Georghiades,A.S., Belhumeur,P.N., Kriegman,D.J. : From Few to Many: Illumination Cone Models for Face Recognition under Variable Lighting and Pose. IEEE Transactions on Pattern Analysis and Machine Intelligence(2001)
- [16] Phillips,P.J., Moon,H., Rauss,P., Rizvi,S.A.: The FERET Evaluation Methodology for Face-Recognition Algorithms. Proceedings of Computer Vision and Pattern Recognition, Puerto Rico, 137-143, 1997
- [17] Kong,H., Xuchun,L., Wang,L., Teoh,E.K., Jian-Gang,W., Venkateswarlu,R. : Generalized 2D Principal Component Analysis. IEEE International Joint Conference on Neural Networks(IJCNN), Montreal, Canada(2005)
- [18] Zhang,D., Zhou,Z., Chen,S. :Diagonal Principal Component Analysis for Face Recognition. Pattern Recognition 39(2006)140–142
- [19] Daniel,J.J., Zia-ur,R., Glenn,A.W. : Properties and Performance of a Center/Surround Retinex. IEEE Transactions on Image Processing, Vol. 6,(1997)
- [20] K.Fuknnaga: Introduction to Statistical Pattern Recognition, Academic Press. second edition(1991)

Discrete Choice Models for Static Facial Expression Recognition

Gianluca Antonini¹, Matteo Sorci¹, Michel Bierlaire², and Jean-Philippe Thiran¹

¹ Ecole Polytechnique Federale de Lausanne, Signal Processing Institute
Ecublens, 1015 Lausanne, Switzerland

{Matteo.Sorci, Gianluca.Antonini, JP.Thiran}@epfl.ch

² Ecole Polytechnique Federale de Lausanne, Operation Research Group
Ecublens, 1015 Lausanne, Switzerland

Michel.Bierlaire@epfl.ch

Abstract. In this paper we propose the use of Discrete Choice Analysis (DCA) for static facial expression classification. Facial expressions are described with expression descriptive units (EDU), consisting in a set of high level features derived from an active appearance model (AAM). The discrete choice model (DCM) is built considering the 6 universal facial expressions plus the neutral one as the set of the available alternatives. Each alternative is described by an utility function, defined as the sum of a linear combination of EDUs and a random term capturing the uncertainty. The utilities provide a measure of likelihood for a combinations of EDUs to represent a certain facial expression. They represent a natural way for the modeler to formalize her prior knowledge on the process. The model parameters are learned through maximum likelihood estimation and classification is performed assigning each test sample to the alternative showing the maximum utility. We compare the performance of the DCM classifier against Linear Discriminant Analysis (LDA), Generalized Discriminant Analysis (GDA), Relevant Component Analysis (RCA) and Support Vector Machine (SVM). Quantitative preliminary results are reported, showing good and encouraging performance of the DCM approach both in terms of recognition rate and discriminatory power.

1 Introduction

Facial expressions are probably the most visual method to convey emotions and one of the most powerful means to relate to each other. An automatic system for the recognition of facial expressions is based on a representation of the expression, learned from a training set of pre-selected meaningful features. For unseen expressions, the corresponding representation has to be associated with the correct expression. In this process, two are the key tasks: the choice of the set of features representing the expression and the choice of the classification rule. In [1] the author focuses on optical flow analysis for feature extraction, in order to model muscle activities and estimating the displacements of salient points. This is a dynamic approach, where temporal information is used both in the feature extraction and classification steps, the last performed through an Hidden Markov Models (HMM) scheme. Gabor wavelet based filters have been used in [2], in order to build templates for facial expressions, over multiple scales and different orientations. Template-based matching is used in order to associate an observed

feature vector with the corresponding expression, in a static context. Statistical generative models such as principal and independent component analysis (PCA, ICA) are used in [3] and [4], in order to capture meaningful statistics of face images. Neural Networks (NN) and HMMs are used for the classification step, respectively in static and dynamic frameworks. Recent years have seen the increasing use of feature geometrical analysis ([5,6]). The Active Appearance Model (AAM, see [7]) is one of these techniques which elegantly combines shape and texture models, in a statistical framework, providing as output a mask of face landmarks.

The contribution of this work is twofold. First, we propose Discrete Choice Models for expression classification. These models have been recently introduced in the computer vision community by [8], in the context of pedestrian modeling and tracking. DCMs are econometric models designed to forecast the behavior of individuals in choice situations, when the set of available alternatives is finite and discrete. In this context, the logic behind the use of DCMs is to model the choice process representing the human observer labelling procedure. The DCM classifier is compared with several other classification methods: LDA, GDA, RCA and SVM. The LDA is a supervised discriminative method to produce the optimal linear classification function. It transforms the data into a lower-dimensional space where it is decided, according to some chosen metric, to which class a given sample x belongs. The GDA is the kernel-based version of the LDA. RCA is a method that seeks to identify and down-scale global unwanted variability within the data. The method performs a projection of the input data into a feature space by means of a linear transformation. In the transformed space, a nearest neighbor classification based on the Euclidean distance is used, in order to assign the new sample to a class (see [9]). Second, we propose a set of *Expression Descriptives Units* (EDU) for static expression representation. They are derived from a set of 55 face landmarks, obtained using an AAM model. The EDUs represent intuitive descriptors of the facial components (eyebrows, eyes, nose and mouth) and the mutual interactions between them. They have been derived taking inspiration from the Facial Action Unit Coding System (FACS) [10] which is a human-observer based system designed to detect subtle changes in facial features. FACS itself is purely descriptive, uses no emotion or other inferential labels and provides the necessary ground-truth with which to describe facial expression. On the other hand, FACS require a huge set of salient facial points, and for most of them a tracking step is required, in order to capture variations over time. EDUs can be considered as a more compact and static counterpart of the FACS.

The paper is structured as follows: in Section 2 we review the AAM and introduce the DCM theory. In Section 3 a detailed description of the utility functions is given along with the EDU description and the results of the learning process. We finally report the experiments and a description of the data used to compare the different classifiers with our approach in Section 4. Conclusions and future works are finally reported in Section 5.

2 Background

2.1 Active Facial Appearance Model

The AAM is a statistical method for matching a combined model of shape and texture to unseen faces. The combination of a model of shape variation with a model of

texture variation generates a statistical appearance model. The model relies on a set of annotated images. A training set of images is annotated by putting a group of landmark points around the main facial features, marked in each example. The shape is represented by a vector \mathbf{s} brought into a common normalized frame -w.r.t. position, scale and rotation- to which all shapes are aligned. After having computed the mean shape $\bar{\mathbf{s}}$ and aligned all the shapes from the training set by means of a Procrustes transformation, it is possible to warp textures from the training set onto the mean shape $\bar{\mathbf{s}}$, in order to obtain shape-free patches. Similarly to the shape, after computing the mean shape-free texture $\bar{\mathbf{g}}$, all the textures in the training set can be normalized with respect to it by scaling and offset of luminance values. PCA is applied to build the statistical shape and textures models:

$$\mathbf{s}_i = \bar{\mathbf{s}} + \Phi_s \mathbf{b}_{si} \quad \text{and} \quad \mathbf{g}_i = \bar{\mathbf{g}} + \Phi_t \mathbf{b}_{ti} \quad (1)$$

where \mathbf{s}_i and \mathbf{g}_i are, respectively, the synthesized shape and shape-free texture, Φ_s and Φ_t are the matrices describing the modes of variation derived from the training set, \mathbf{b}_{si} and \mathbf{b}_{ti} the vectors controlling the synthesized shape and shape-free texture. The unification of the presented shape and texture models into one complete appearance model is obtained by concatenating the vectors \mathbf{b}_{si} and \mathbf{b}_{ti} and learning the correlations between them by means of a further PCA. The statistical model is then given by:

$$\mathbf{s}_i = \bar{\mathbf{s}} + Q_s \mathbf{c}_i \quad \text{and} \quad \mathbf{g}_i = \bar{\mathbf{g}} + Q_t \mathbf{c}_i \quad (2)$$

where Q_s and Q_t are the matrices describing the principal modes of the combined variations in the training set and \mathbf{c}_i is the appearance parameters vector, allowing to control simultaneously both shape and texture. Fixing the parameters \mathbf{c}_i we derive the shape and the shape-free texture vectors using equations (2). A full reconstruction is given by warping the generated texture into the generated shape. In order to allow pose displacement of the model, other parameters must be added to the appearance parameters \mathbf{c}_i : the pose parameters \mathbf{p}_i . The matching of the appearance model to a target face can be treated as an optimization problem, minimizing the difference between the synthesized model image and the target face [7].

2.2 Discrete Choice Models

Discrete choice models are known in econometrics since the late 50's. They are defined to describe the behavior of people in choice situations, when the set of available alternatives is finite and discrete (choice set). They are based on the concept of *utility maximization* in economics, where the decision maker is assumed to be *rational*, performing a choice in order to maximize the utilities she perceives from the alternatives. The alternatives are supposed to be mutually exclusive and collectively exhaustive, while the rationality of the decision maker implies transitive and coherent preferences.¹ The utility is a *latent* construct, which is not directly observed by the modeler, and is treated as

¹ Transitive preferences means that if alternative i is preferred to alternative j which is preferred to alternative k , then alternative i is also preferred to k . Coherent preferences means that the decision maker will make the same choice in exactly the same conditions.

a random variable. The discrete choice paradigm well matches the labelling assignment process in a classification task. This approach can be interpreted as an attempt to model the decision process performed by an hypothetical human observer during the labelling procedure for the facial expressions.

Given a population of N individuals, the (random) utility function U_{in} perceived by individual n from alternative i , given a choice set C_n , is defined as follows:

$$U_{in} = V_{in} + \varepsilon_{in} \tag{3}$$

with $i = 1, \dots, J$ and $n = 1, \dots, N$. V_{in} represents the deterministic part of the utility, which is a function of alternatives' attributes and socio-economic characteristics of the decision maker. In the context of this paper, we only deal with attributes of the alternatives, represented by combinations of the chosen features. The ε_{in} term is a random variable capturing the uncertainty. Under the utility maximization assumption, the output of the model is represented by the choice probability that individual n will choose alternative i , given the choice set C_n . It is given by:

$$P_n(i|C_n) = P_n(U_{in} \geq U_{jn}, \forall j \in C_n, j \neq i) = \int_{\varepsilon_n} I(\varepsilon_n < V_{in} - V_{jn}, \forall j \in C_n, j \neq i) f(\varepsilon_n) d\varepsilon_n \tag{4}$$

where $\varepsilon_n = \varepsilon_{jn} - \varepsilon_{in}$ and $I(\cdot)$ is an indicator function which is equal to 1 when its argument is satisfied, zero otherwise. Based on Equation 4, in order to define the choice probability, only the difference between the utilities matters. The specification of the utility functions represents the modeler's mean to add her prior knowledge on the choice process (a similar interpretation of the decision theoretic approach can be found in [11]). In this sense, the DCM approach is similar to graphical probabilistic models, such as belief networks and random fields, where the graph topology embeds the prior knowledge, helping designing causal relationships. Different DCMs are obtained making different assumptions on the error terms. A family of models widely used in literature are the GEV (Generalized Extreme Value) models, introduced by [12]. GEV models provide a closed form solution for the choice probability integral, allowing at the same time for a certain flexibility in designing the variance/covariance structure of the problem at hand (i.e., several correlation patterns between the alternatives can be explicitly captured by these models). Assuming the error terms being multivariate type I extreme value distributed², the general expression of the GEV choice probability for a given individual to choose alternative i , given a choice set C with J alternatives, is as follows:

$$P(i|C) = \frac{e^{V_i + \log G_i(y_1, \dots, y_J)}}{\sum_{j=1}^J e^{V_j + \log G_j(y_1, \dots, y_J)}} \tag{5}$$

where $y_i = e^{V_i}$ and $G_i = \frac{\partial G}{\partial y_i}$. The function G is called *generating function* and it captures the correlation patterns between the alternatives. Details about the mathematical

² The main reasons for the choice of this kind of distribution derive from its good analytical properties. More details can be found in [13].

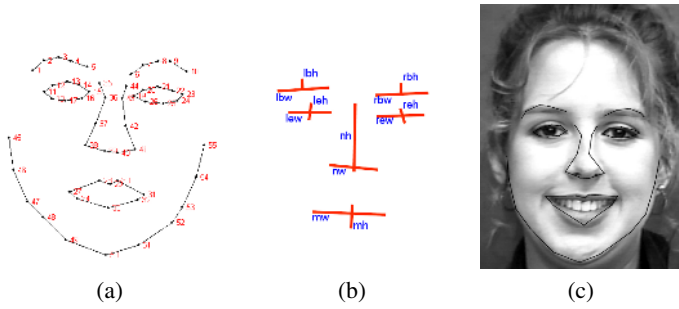


Fig. 1. a)Facial landmarks (55 points);b)Facial components descriptors;c)Expressions Descriptive Units

properties of G are reported in [12] (differentiable and homogeneous of degree $\mu > 0$, among the others). Several GEV models can be derived from Equation 5, through different specifications of the generating function. In this paper we use a Multinomial Logit Model (MNL), which is largely the simplest and most used discrete choice model in literature. It is obtained assuming the following G function, which implies no correlations between the alternatives:

$$G(y_1, \dots, y_J) = \sum_{j \in C} y_j^\mu \tag{6}$$

where μ is a positive scale parameter. Under these assumptions, the MNL choice probability is given by the following expression

$$P(i|C) = \frac{e^{\mu V_i}}{\sum_{j \in C} e^{\mu V_j}} \tag{7}$$

3 DCM and Facial Expression Classification

3.1 Expressions Descriptive Units

The use of AAM allows to detect facial components in a face (an example is shown in Figure 1(c)). Figure 1(a) shows the 55 landmarks used to build the AAM model,

Table 1. Expressions Descriptive Units

EDU1	$\frac{lew+rew}{leh+reh}$	EDU8	$\frac{leh+reh}{lbh+rbh}$
EDU2	$\frac{lbw}{lbh}$	EDU9	$\frac{lew}{nw}$
EDU3	$\frac{rbw}{rbh}$	EDU10	$\frac{nw}{mw}$
EDU4	$\frac{mw}{mh}$	EDU11	EDU2 / EDU4
EDU5	$\frac{nh}{nw}$	EDU12	EDU3 / EDU4
EDU6	$\frac{lew}{mw}$	EDU13	EDU2 / EDU10
EDU7	$\frac{leh}{mh}$	EDU14	EDU3 / EDU10

Table 2. Utility functions: each row corresponds to an expression while the columns are the EDUs included in the utilities

Expressions	edu1	edu2	edu4	edu5	edu6	edu7	edu8	edu10	edu11	edu13	edu14
anger	✓			✓	✓					✓	✓
disgust	✓			✓					✓	✓	✓
fear			✓	✓	✓			✓			✓
happiness	✓		✓	✓			✓	✓			
neutral			✓								
sadness		✓					✓				
surprise			✓			✓					

while Figure1(b) shows the descriptors we use for the facial components: eyes, eye-brows, nose and mouth. These descriptors represent the width and the height of each facial component. In order to give a useful representation of the expression in terms of interactions among those descriptors, we define a set of EDUs, reported in Table 1. The first 5 EDUs represent, respectively, the eccentricity of eyes, left and right eyebrows, mouth and nose. The EDUs from 7 to 9 represent the eyes interactions with mouth and nose, while the 10th EDU is the nose-mouth relational unit. The last 4 EDUs relate the eyebrows to mouth and nose. Differently from other approaches [14,9] that use the combined AAM vector parameters as facial features, in our framework the 14 EDUs represent the features describing the face. The intuitive interpretation and the reduced number of dimensions make of the EDUs a valid set of descriptors for facial expressions.

3.2 The Model

The utility functions are specified using a linear-in-parameters form, combining the expression descriptive units. Each EDU in each utility is weighted by an unknown deterministic coefficient, that has to be estimated. The choice for a linear form is based purely on simplicity considerations, in order to reduce the number of parameters in the estimation process. The general form of the utilities is given by:

$$U_i = \alpha_i + \sum_{k=1}^K I_{ki} \beta_{ki} \text{EDU}_k \tag{8}$$

where $i = 1, \dots, C$ with $C = 7$ is the number of expressions, $K = 14$ is the number of EDUs, I_{ki} is an indicator function equal to 1 if the k -th EDU is included in the utility for expression i and 0 otherwise, β_{ki} is the weight for the k -th EDU in alternative i and α_i is an alternative specific constant. The α_i coefficients represent the average value of the unobserved part of the corresponding utility and one of them has to be normalized to 0, in order to be consistent with DCM theory (see [13]). In our case, we normalize with respect to the neutral expression. We summarize in Table 2 what are the EDUs included in the different utilities, i.e. when the $I_{ki} = 1$. Table 2 shows how, during the model specification step, we are free to customize the utilities of the different expressions. This flexibility represents the strength of DCMs; note that the utility expressions reported

here are the result of a strong iterative process, where several hypothesis have been tested and validated, starting from a uniform expression for every alternative, including all the EDUs. In the final utility functions, only the EDUs corresponding to statistically significant parameters (t -test statistic against the zero value) are reported, resulting in a final model with 31 unknown parameters (6 α_i and 25 β_{ki}). The parameters have been estimated by maximum likelihood estimation, using the Biogeme package [15]. Biogeme is a freeware, open source package available from roso.epfl.ch/biogeme. It performs maximum likelihood estimation and simulated maximum likelihood estimation of a wide class of random utility models, within the class of mixtures of Generalized Extreme Value models (see [16] for details). The maximization is performed using the CFSQP algorithm (see [17]), using a Sequential Quadratic Programming method. Note that such nonlinear programming algorithms identify local maxima of the likelihood function. We performed various runs, with different starting points (a trivial model with all parameters to zero, and the estimated value of several intermediary models). They all converged to the same solution. Most of the estimated utility parameters are significantly different from zero. Classification is performed running the learned model on the test set, using the BioSim package (available at the same address as Biogeme). BioSim performs a sample enumeration on the test data, providing for each of them the utilities and the choice probabilities for each expression in the choice set. The classification rule consists in associating each sample with the alternative having the maximum probability, whose equation is reported in (7). We can state the classification rule used here as a *soft max* principle based on an *entropy maximization* criterion ([18]). However, such a (only formal) 'equivalence' arises on the base of the specific form of the GEV probability equation. Other discrete choice methods exist (Probit, Logit Kernel, [16,19]) whose choice probabilities cannot be expressed by an analytical solution, leading to a more general soft max classification scheme, not related with the maximum entropy principle.

Learning results. The learned parameters show important consistencies with the common reading of facial expressions in terms of facial component modifications. For space reasons, we report in Table 3 only a subset of β_{ki} estimates.

The parameters β_{5a} represents the coefficient of the 5th EDU (nose eccentricity) for the anger alternative. Its positive value shows a positive impact on the respective utility. It means that increasing nose eccentricity corresponds to higher utilities for the anger alternative. Looking at the definition of this EDU, this is in line with our expectations, showing that for an anger expression the nose width increases while its height decreases, with respect to the neutral expression (the reference one in our model). The parameters β_{1a} represents the eye eccentricity (1st EDU) for the anger expression. A similar interpretation holds for this coefficient, in line with observations: the eye movement leads to a lower eye's height and a higher eye's width, with respect to the reference alternative. The other two parameters relate the nostrils width with the mouth width. Their negative sign induces a negative impact on the utilities of fear and happiness. This is coherent with the data, where for these two expressions we note a characterizing increase in the mouth width, leading to a decreasing nostril-mouth interaction parameter. The coefficient estimates are significantly different from zero at 95%, with the exception of the β_{10f} (significant at 90%). We finally report some interesting statistics. The

Table 3. MNL Part of the estimation results

β_{ki}	estimate	t test 0
β_{5a}	+ 1.238	+ 4.298
β_{1a}	+ 2.067	+ 2.018
β_{10f}	- 14.69	- 1.871
β_{10h}	- 42.64	- 3.440
Sample size = 143		
Number of estimated parameters = 30		
Null log-likelihood = - 278.265		
Final log-likelihood = - 88.317		
Likelihood ratio test = 379.896		
$\bar{\rho}^2 = 0.575$		

Table 4. Number of images in the classification training and test set

Expressions	Training images	Test images
Neutral	26	15
Happiness	20	18
Surprise	21	20
Fear	18	11
Anger	18	17
Disgust	22	17
Sadness	18	17

Table 5. Classification rates

Classifiers	Classification Rate(%)
DCM	78.261
SVM	76.522
RCA	70.435
LDA	49.565
GDA	62.609

log-likelihood corresponding to a trivial model (all the coefficients equal to zero) is consistently increased after the estimation process, rising its value from -278.265 to -88.317. The likelihood ratio test and the $\bar{\rho}^2$ coefficient are also reported, showing the good fitting of the estimated model.

4 Experiments

In order to test the proposed approach we use the Cohn-Kanade Database [20]. The database consists of expression sequences of subjects, starting from a neutral expression and ending most of the time in the peak of the facial expression. There are 104 subjects in the database, but only for few of them the six expressions are available. From the database we extrapolate 3 data sets:

- *AAM training set*: it consists of 300 images from 11 different subjects; it is composed by 48 neutral images and 42 images for each of the 6 primary emotions.
- *Classifiers training and test set*: they consist respectively of 143 and 115 appearance masks, as reported in Table 4.

Table 6. DCM and SVM confusion matrices

DCM/SVM	happiness	surprise	fear	anger	disgust	sadness	neutral	Overall(%)
happiness	16/16	0/0	2/1	0/0	0/1	0/0	0/0	88.88/88.88
surprise	0/0	19/19	0/0	1/0	0/0	0/0	0/1	95.00/95.00
fear	1/4	0/0	7/4	2/2	0/0	1/0	0/1	63.64/36.36
anger	0/0	0/0	1/2	10/9	3/1	2/3	1/2	58.82/52.94
disgust	0/1	0/0	0/0	2/2	12/12	0/1	3/1	70.58/70.58
sadness	0/0	0/0	0/0	2/0	0/1	15/14	0/2	88.24/82.35
neutral	0/1	0/0	2/0	1/0	0/0	1/0	11/14	73.34/93.33

The appearance model is built using 49 shapes modes and 140 texture modes leading to 84 appearance modes, capturing the 98% of the combined shape and texture variation. The shape-free texture vector \mathbf{g} is composed of 38310 pixels and the shape vector dimension is 55. Concerning the implementation, we use the AAM C++ code available at <http://www2.imm.dtu.dk/~aam/>. The classifiers used in the comparison procedure all share the same training and test sets. Their input consists in the EDUs built on the matched appearance masks. For the other classifiers we implemented the related state of the art. The SVM (using `libsvm` with radial basis functions) and DCM classifiers have been tuned on the common training set. The test experiments reported in Table 5, although preliminaries, could be interpreted as a better transferability of the learned DCM model over unseen samples. The intuition explaining this behaviour could lie on the more flexible hypotheses at the base of the DCM approach. The verification of this intuition will be part of our further investigations.

With the exception of RCA, the performance of the other nearest neighbor classifiers are significantly lower than SVM and DCM. For this reason we report in Table 6 only the confusion matrices for the two best performing methods.

A second empiric measure for classifiers performance comparison, related only to their discriminatory power and not to the recognition rates, has been computed. It is described in [21] and we report here a short explanation. For the various methods to be compared, let $m(wc)$ the mean probability assigned to well classified samples and $std(wc)$ the relative standard deviation. Similarly, let $m(bc)$ and $std(bc)$ the same values for the bad classified samples.³ Good classification and bad classification thresholds are defined as:

$$gctr = m(wc) + std(wc) \quad bctr = m(bc) - std(bc)$$

Based on these values, an overall performance parameter is defined as:

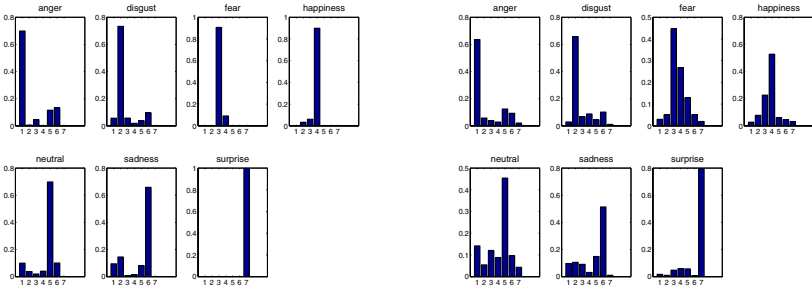
$$opp = \frac{gctr - bctr}{m(bc) - m(wc)} \quad (9)$$

measuring how well a classifier discriminates. For a robust method we expect an opp value as low as possible. If the value of opp is negative, the $gctr$ and $bctr$ thresholds are well separated. In Table 7 we report the opp values for the tested classifiers. In

³ For the nearest neighbor classifiers the sample distances from the classes have been normalized, in order to sum up to one.

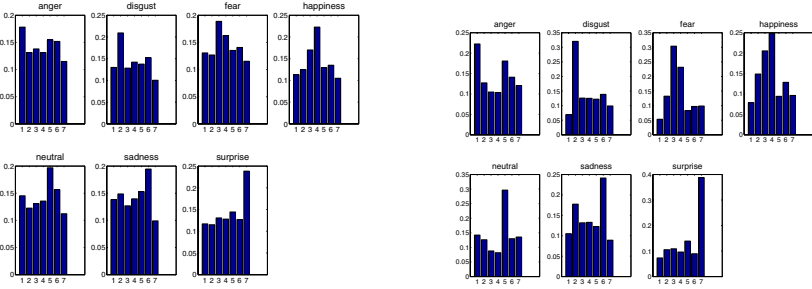
Table 7. opp values for the compared classifiers

Classifiers	opp
DCM	- 6.92
SVM	- 3.60
RCA	- 3.64
LDA	- 4.51
GDA	- 3.27



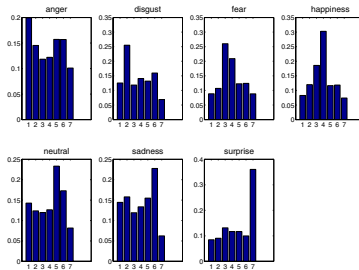
(a) DCM

(b) SVM



(c) RCA

(d) LDA



(e) GDA

Fig. 2. Mean class probabilities for correct classification

order to visualize the discriminatory power for DCM and SVM, it is worth to show, for each set of correctly classified samples, the mean values of the probabilities assigned to each of the other classes, as shown in Figure 2. The plots confirm the better class discrimination performed by the DCM classifier, resulting in a sharper shape of the output probabilities.

5 Conclusions and Future Works

In this paper we propose a new classifier based on discrete choice analysis and a set of expression descriptive units for facial expression representation. Both the feature set and the modeling approach are motivated by the research of methods able to bring into the process the modeler prior knowledge. The set of the proposed EDUs is suitable because intuitively related to static expressions, describing the salient facial components and their mutual interactions. The DCM modeling approach redefines facial expression classification as a discrete choice process, which well matches the human observer labelling procedure. Prior knowledge can be included in the process customizing the utilities. The result of the DCM is a set of probabilities assigned to the alternatives, represented by the possible expressions. The one with the maximum probability is chosen for classification. We compared the DCM classifier with several other methods, finding that only the SVM has comparable performance. However, the more flexible properties of DCM lead to better results of this approach both in terms of classification rate on new data and discriminatory power. We are currently working to include in the model both the expression dynamics and the variation in a population of individuals performing the labelling task. Based on our experience, we think that a subjective component biases the labelling process, requiring a detailed statistical analysis on collected data from an heterogeneous population of human observers. DCMs, coming from econometric, provide a strong statistical framework to include such a heterogeneity.

References

1. Lien, J.: Automatic recognition of facial expressions using hidden markov models and estimation of expression intensity (1998)
2. Ye, J., Zhan, Y., Song, S.: Facial expression features extraction based on gabor wavelet transformation. In: IEEE International Conference on Systems, Man and Cybernetics. (2004) 10–13
3. Padgett, C., Cottrell, G.: Representing face images for emotion classification. MIT Press, Cambridge, MA (1997)
4. Bartlett, M.: Face image analysis by unsupervised learning and redundancy reduction (1998)
5. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models. *Pattern Analysis and Machine Intelligence, IEEE Transactions on* **23** (2001) 681–685
6. Lanitis, A., Taylor, C.J., Cootes, T.F.: Automatic interpretation and coding of face images using flexible models. *IEEE Trans. Pattern Anal. Mach. Intell.* **19** (1997) 743–756
7. Stegmann, M.B.: Active appearance models: Theory, extensions and cases. Master's thesis, Informatics and Mathematical Modelling, Technical University of Denmark, DTU, Richard Petersens Plads, Building 321, DK-2800 Kgs. Lyngby (2000)

8. Antonini, G., Venegas, S., Bierlaire, M., Thiran, J.P.: Behavioral priors for detection and tracking of pedestrians in video sequences. To appear in *International Journal of Computer Vision* (2005)
9. Sorci, M., Antonini, G., Thiran, J.P.: Relevant component analysis for static facial expression recognition. Technical Report TR_ITS_2005.33, Signal Processing Institute, Ecole Polytechnique Federale de Lausanne (2005)
10. Ekman, P., Friesen, W.V.: *Facial Action Coding System Investigator's Guide*. Consulting Psychologist Press, Palo Alto, CA (1978)
11. Horvitz, E.J., Breese, J.S., Henrion, M.: Decision theory in expert systems and artificial intelligence. *International Journal of Approximate Reasoning* **2** (1988) 247–302
12. McFadden, D.: Modelling the choice of residential location. In A. Karlquist *et al.*, ed.: *Spatial interaction theory and residential location*, Amsterdam, North-Holland (1978) 75–96
13. Ben-Akiva, M.E., Lerman, S.R.: *Discrete Choice Analysis: Theory and Application to Travel Demand*. MIT Press, Cambridge, Ma. (1985)
14. Abboud, D., Davoine, F.: Appearance factorization based facial expression recognition and synthesis. In: *ICPR* (4). (2004) 163–166
15. Bierlaire, M.: BIOGEME: a free package for the estimation of discrete choice models. In: *Proceedings of the 3rd Swiss Transportation Research Conference, Ascona, Switzerland* (2003) www.strc.ch.
16. Train, K.: *Discrete Choice Methods with Simulation*. Cambridge University Press, University of California, Berkeley (2003)
17. Lawrence, C.T., Zhou, J.L., Tits, A.: A c code for solving (large scale) constrained nonlinear (minimax) optimization problems, generating iterates satisfying all inequality constraints. Technical Report TR-94-16rl, Institute for Systems Research, University of Maryland, College Park, MD 20742 (1997)
18. Cover, T.M., Thomas, J.A.: *Elements of information theory*. Wiley (1991)
19. Ben-Akiva, M., Bolduc, D.: Multinomial probit with a logit kernel and a general parametric specification of the covariance structure (1996) Working Paper, Department of Civil Engineering, MIT.
20. Kanade, T., Cohn, J., Tian, Y.L.: Comprehensive database for facial expression analysis. In: *Proceedings of the 4th IEEE International Conference on Automatic Face and Gesture Recognition (FG'00)*. (2000) 46 – 53
21. Sujith, K.R., Ramanan, G.V.: Procrustes analysis and moore-penrose inverse based classifiers for face recognition. In: *IWBRIS*. (2005) 59–66

Scalable and Channel-Adaptive Unequal Error Protection of Images with LDPC Codes

Adrian Munteanu, Maryse R. Stoufs, Jan Cornelis, and Peter Schelkens

Vrije Universiteit Brussel, Department of Electronics and Informatics,
Pleinlaan 2, 1050 Brussel, Belgium
{acmuntea, mstoufs, jpcornel, pschelke}@etro.vub.ac.be

Abstract. This paper considers the design of an optimal joint source-channel coding system employing scalable wavelet-based source coders and unequal error protection for error-resilient transmission over binary erasure channels. We theoretically show that the expected average rate-distortion function in the separately encoded subbands is convex with monotonically decreasing slopes and use this fact to propose a new algorithm which minimizes the end-to-end distortion. This algorithm is based on a simplified Viterbi-search followed by Lagrangian-optimization. We show that our proposed solution results in significant complexity reductions while providing very near-to-optimal performance.

1 Introduction

The transmission of images over heterogeneous, error-prone networks to a large variety of terminals requires scalable image coding and efficient error control mechanisms. Many of the state-of-the-art scalable image coders rely on a global wavelet-transform followed by embedded bitplane coding and entropy coding [1], [2], [3]. These source coders are attractive because they offer very competitive rate-distortion performances together with resolution and quality scalability, such that progressive transmission and bandwidth adaptation can be realized with one single encoded bit stream. However, the nature of the entropy coding techniques used in these image coding schemes, causes the produced bitstreams to be very sensitive to transmission errors; i.e. even a single bit-error may propagate and cause the decoder to lose synchronization and eventually collapse. Error control coding is therefore of vital importance when transmitting the generated bit streams over error-prone channels.

Error protection in packet-based systems is typically enabled by means of forward error correction (FEC) codes. The last decade has been marked with exciting developments in this field with the introduction of iteratively decodable and capacity-achieving codes like turbo codes in 1993 by Berrou and Glavieux [4] and the independent re-discovery of the low-density parity-check (LDPC) codes, originally introduced by Gallager in the early 1960's [5], by MacKay [6] in 1995. Driven by these results other similar codes have recently been proposed [7], [8], [9].

Shannon's separation theorem [10] states that source and channel coding can be performed independently, while maintaining optimality. However, this is only true under the assumption of asymptotically long block lengths of data and unlimited complexity and delay. In many applications, these conditions neither hold, nor can they be

used as a good approximation. It has been shown in the past [11] that a joint design, allowing for the optimal allocation of the available bits between the source and channel codes, leads to significantly improved coding results.

In this paper we focus on the joint source and channel coding (JSCC) problem, wherein the source is encoded by a scalable codec producing layers with different levels of importance. We will follow a JSCC-design that accounts for the scalability of the source coder by incorporating unequal error protection (UEP) [12] of the source packets. This leads to a better performance than an equivalent JSCC-design which performs equal error protection [13]. In the context of the packet-based transmission of embedded sources, JSCC-algorithms with UEP were recently proposed in [13], [14], [15] for the case of fixed-size channel-packets and in [13], [15] for the case of variable-length channel-packets. In [14], a distortion-based rate-allocation method was presented for protecting JPEG2000 bitstreams with turbo codes. It employs a complete list Viterbi-search and its complexity grows quadratically with the number of transmitted packets. The performance of the algorithm approaches the one obtained by using exhaustive search. A Lagrangian optimization-based JSCC-approach that iteratively computes the source rates and protection levels for a JPEG2000-bitstream is designed in [13]. In [15] a distortion-optimal solution that first performs a sub-optimal rate-based search [16] followed by a distortion-based local search is proposed.

In this paper, we propose a Lagrangian optimization-based JSCC-algorithm with UEP which minimizes the end-to-end distortion. In contrast to [13], the proposed algorithm constructs unique optimal rate-distortion curves for each code-block (subband), and based on them, performs optimum joint source-channel rate-allocation. Unlike the algorithms described in [13], [14], [15], we focus on transmission over packet-erasure channels instead of binary symmetric channels. The erasure channel is a better model for modern packet-based networks, since packets either arrive at the destination or get lost due to congestion or corruption [17]. The proposed JSCC approach assumes that an interleaver is used in the transmission scheme, so that the packet-loss model can be translated into a binary-erasure channel (BEC) model. The coding system employs punctured regular LDPC-codes and fixed-length channel packets, since this provides the advantage of an easier cross-layer design [18].

This paper is structured as follows. Section 2 formulates the problem under investigation as a constrained-optimization problem. Section 3 presents our solution and gives the complete derivation of our JSCC-methodology with UEP. Section 4 highlights the specifics of our source and channel coder. Section 5 reports the simulation results. Finally, section 6 draws the conclusions of our proposed methodology.

2 Problem Formulation

We consider an image decomposed into L wavelet subbands which are progressively encoded and focus on the transmission of fixed-length packets over a binary erasure channel with parameter ε and total channel rate R_{tot} (capacity of the channel). The JSCC has to allocate the total rate R_{tot} across all L subbands and between the source and channel coders in such a way that the overall distortion is minimized.

The total expected distortion of the transmitted image can be expressed as a sum of the separate subband-distortions \overline{D}_l as: $\overline{D}_{tot} = \sum_{l=1}^L \overline{D}_l(R_{s,l}, R_{c,l})$. The problem is to find for every subband l the optimum source and channel rates $R_{s,l}$ and $R_{c,l}$ respectively as well as the corresponding number of packets M_l . Additionally, for every given subband, the optimal rate distribution between the source and channel coders at the level of every packet needs to be determined. Denote by $R_{s_i,l}, R_{c_i,l}$ the source and channel rates respectively used in packet i of subband l . The global rate is given by:

$$R_{tot} = \sum_{l=1}^L w_l R_l = \sum_{l=1}^L w_l (R_{s,l} + R_{c,l}) = \sum_{l=1}^L w_l \left(\sum_{i=0}^{M_l} R_{s_i,l} + \sum_{i=0}^{M_l} R_{c_i,l} \right). \tag{1}$$

The weights w_l reflect the subband contribution in the total rate, which is proportional to the relative subband size, i.e. $w_l = S_l/S$, where S_l, S are the areas of the subband and original image respectively. In our approach, we define a number of protection levels d_{tot} and impose the constraint of UEP, i.e. the source rate in the M_l packets has to be non-decreasing: $R_{s_1,l} \leq R_{s_2,l} \leq \dots \leq R_{s_{M_l},l}$.

The JSCC problem can thus be formulated as a constrained-optimization problem wherein one has to minimize the expected distortion:

$$\min \left(\overline{D}_{tot} = \sum_{l=1}^L \overline{D}_l(R_{s,l}, R_{c,l}) \right) \tag{2}$$

subject to the constraint that the target rate is closely met:

$$R_{tot} = \sum_{l=1}^L w_l R_l = \sum_{l=1}^L w_l (R_{s,l} + R_{c,l}) \leq R_{target}. \tag{3}$$

3 Solution Methodology

The proposed JSCC approach is presented next. We first derive a recursive formula for the average subband distortion. Thereafter, we prove that the JSCC solution can be found via Lagrangian optimization. We also show that it is computationally impossible to create all possible protection combinations for all subbands and propose a novel strategy to find a close-to-optimal solution for the constrained-optimization problem.

3.1 Recursive Formula for the Average Expected Subband-Distortion

Denote by $p_f(\epsilon)$ the probability of losing a packet that is transmitted over a BEC with parameter ϵ . The expected subband-distortion $\tilde{D}_l(R_{s,l}, R_{c,l})$ when receiving all packets up to packet m (with $0 \leq m \leq M_l$) and losing packet $m+1$ is of the form:

$$\tilde{D}_l(R_{s,l}, R_{c,l}) = \tilde{D}_l \left(\sum_{i=0}^m R_{s_i,l}, \sum_{i=0}^m R_{c_i,l} \right) = \prod_{i=0}^m (1 - p_{f_i}(\epsilon)) \cdot p_{f_{m+1}}(\epsilon) \cdot D_l \left(\sum_{i=0}^m R_{s_i,l} \right), \tag{4}$$

where D_l is the source distortion, $p_{f_0} = 0$, $p_{f_{M_l+1}}(\epsilon) = 1$ and $R_{s_0,l} = R_{c_0,l} = 0$. Hence, the average expected distortion of subband l when transmitting M_l packets is:

$$\overline{D}_l(R_{s,l}, R_{c,l}) = \sum_{m=0}^{M_l} \tilde{D}_l \left(\sum_{i=0}^m R_{s_i,l}, \sum_{i=0}^m R_{c_i,l} \right) = \sum_{m=0}^{M_l} \prod_{i=0}^m (1 - p_{f_i}(\epsilon)) \cdot p_{f_{m+1}}(\epsilon) \cdot D_l \left(\sum_{i=0}^m R_{s_i,l} \right). \tag{5}$$

Next, we define the code rate r of the error correction codes as:

$$r \triangleq \frac{k}{N} \quad (6)$$

where k is the number of source bits and N is the total number of bits in the channel packet. In our transmission scenario, N is fixed. Hence, $D_l(\sum_{i=0}^m R_{s_i,l})$ can equivalently be formulated as a function of the set of code rates $r_{i,i}$ used in packet i , $0 \leq i \leq M_l$, of subband l as $D_l(\sum_{i=0}^m r_{i,i})$. Equivalently, $\overline{D}_l(R_{s,l}, R_{c,l})$ can be expressed as $\overline{D}_l(r_{l,0}, r_{l,1}, \dots, r_{l,M_l})$. Therefore, (5) can be written as:

$$\overline{D}_l(r_{l,0}, r_{l,1}, \dots, r_{l,M_l}) = \sum_{m=0}^{M_l} \left[\prod_{i=0}^m (1 - p_f(r_{l,i}, \varepsilon)) \right] \cdot p_f(r_{l,m+1}, \varepsilon) \cdot D_l(r_{l,0} + r_{l,1} + \dots + r_{l,m}) \quad (7)$$

where $p_f(r_{l,i}, \varepsilon) \equiv p_{f_i}(\varepsilon)$. We denote: $\alpha_{l,m} = \prod_{i=0}^m (1 - p(r_{l,i}, \varepsilon))$ and $\widetilde{r}_{l,k} = \sum_{i=0}^k r_{l,m}$. Together with the conventions $\alpha_{l,0} = 1$, $r_{l,0} = 0$ and $p(r_{l,M_l+1}) = 1$, we can write:

$$\alpha_{l,m+1} = \alpha_{l,m} \cdot (1 - p(r_{l,m+1}, \varepsilon)) \Leftrightarrow \alpha_{l,m} \cdot p(r_{l,m+1}, \varepsilon) = (\alpha_{l,m} - \alpha_{l,m+1}), \quad (8)$$

for $m, 1 \leq m < M_l$. Using (7) and (8), a *recursive* formula can now easily be derived:

$$\begin{aligned} \overline{D}_l(r_{l,0}, r_{l,1}, \dots, r_{l,M_l}) &= \left[\sum_{m=0}^{M_l-1} (\alpha_{l,m} - \alpha_{l,m+1}) \cdot D_l(\widetilde{r}_{l,m}) \right] + \alpha_{l,M_l} \cdot D_l(\widetilde{r}_{l,M_l}) \\ \overline{D}_l(r_{l,0}, r_{l,1}, \dots, r_{l,M_l-1}) &= \left[\sum_{m=0}^{M_l-2} (\alpha_{l,m} - \alpha_{l,m+1}) \cdot D_l(\widetilde{r}_{l,m}) \right] + \alpha_{l,M_l-1} \cdot D_l(\widetilde{r}_{l,M_l-1}) \\ \Leftrightarrow \overline{D}_l(r_{l,0}, r_{l,1}, \dots, r_{l,M_l}) &= \overline{D}_l(r_{l,0}, r_{l,1}, \dots, r_{l,M_l-1}) + \alpha_{l,M_l} \cdot (D_l(\widetilde{r}_{l,M_l}) - D_l(\widetilde{r}_{l,M_l-1})) \end{aligned} \quad (9)$$

The code rates $(r_{l,0}, r_{l,1}, \dots, r_{l,M_l})$ assigned to the M_l packets of subband l have to be chosen such that a minimal end-to-end distortion is achieved. We call the set of code rates $(r_{l,0}, r_{l,1}, \dots, r_{l,M_l})$ the *path* Π_{M_l} . Our final *recursive formula* is then:

$$\begin{aligned} \overline{D}_l(\Pi_{M_l}) &= \overline{D}_l(\Pi_{M_l-1}, r_{l,M_l}) = \\ &= \overline{D}_l(\Pi_{M_l-1}) + \alpha_{l,M_l-1} \cdot (1 - p_f(\varepsilon, r_{l,M_l})) \cdot (D_l(\widetilde{r}_{l,M_l}) - D_l(\widetilde{r}_{l,M_l-1})) \end{aligned} \quad (10)$$

3.2 Optimal Rate-Allocation with Lagrangian Optimization

In the following we prove that the average expected distortion $\overline{D}_l(\Pi_{M_l})$ of a subband is always convex with monotonically decreasing slopes, no matter which path is taken. Also, we prove that a similar conclusion with bounds on the allowable code rates can be drawn when transmitting an increasing number of fixed-length packets.

Lemma 1: Define $\lambda_{\overline{D}_l}(\Pi_k) = \frac{\overline{D}_l(\Pi_{k-1}) - \overline{D}_l(\Pi_k)}{r_{l,k}}$. If $D_l(\widetilde{r}_{l,k})$ is convex with monotonically decreasing slopes, then $\lambda_{\overline{D}_l}(\Pi_{k-1}) > \lambda_{\overline{D}_l}(\Pi_k)$ for any $k, 1 \leq k \leq M_l$.

Proof:

Using our recursive formula (10) we can state:

$$\lambda_{\overline{D}_l}(\Pi_k) = \frac{\overline{D}_l(\Pi_{k-1}) - \overline{D}_l(\Pi_k)}{r_{l,k}} = \alpha_{l,k-1} \cdot (1 - p_f(\varepsilon, r_{l,k})) \cdot \frac{[D_l(\widetilde{r}_{l,k-1}) - D_l(\widetilde{r}_{l,k})]}{\widetilde{r}_{l,k} - \widetilde{r}_{l,k-1}}$$

$$\Leftrightarrow \lambda_{\overline{D}_l}(\Pi_k) = \alpha_{l,k-1} \cdot (1 - p_f(\varepsilon, r_{l,k})) \cdot \lambda_{D_l}(\widetilde{r}_{l,k}) \Leftrightarrow \lambda_{\overline{D}_l}(\Pi_k) = \alpha_{l,k} \cdot \lambda_{D_l}(\widetilde{r}_{l,k})$$

Since $1 - p_f(\varepsilon, r_{l,k}) \leq 1, \forall r_{l,k}$, it implies $\alpha_{l,k} = \alpha_{l,k-1} \cdot (1 - p_f(\varepsilon, r_{l,k})) \leq \alpha_{l,k-1}$ and because $D_l(\widetilde{r}_{l,k})$ is convex with monotonically decreasing slopes: $\lambda_{D_l}(\widetilde{r}_{l,k}) = \lambda_{D_l}(\widetilde{r}_{l,k-1} + r_{l,k}) < \lambda_{D_l}(\widetilde{r}_{l,k-1})$. We obtain that $\lambda_{\overline{D}_l}(\Pi_{k-1}) > \lambda_{\overline{D}_l}(\Pi_k)$. Q.E.D.

In our transmission scheme we assume that the source is distributed in fixed-length packets. The next lemma derives the sufficient condition on the code rates such that the distortion is monotonically decreasing with the number of transmitted packets.

Lemma 2: Define $\gamma_{\overline{D}_l}(\Pi_k) = \frac{\overline{D}(\Pi_{k-1}) - \overline{D}(\Pi_k)}{k - (k-1)}$, which is the slope of the distortion

when sending packet k after packet $k-1$. Assume that $D_l(\widetilde{r}_{l,k})$ is convex and of the form $D_l(\widetilde{r}_{l,k}) = \varepsilon \sigma^2 2^{-2\widetilde{r}_{l,k}}$. A sufficient condition for $\gamma_{\overline{D}_l}(\Pi_k)$ to be monotonically decreasing with k is $r_{l,k} > \log_4 e/e \approx 0.2654$, for all $k, 1 \leq k \leq M_l$.

Proof:

We observe that $\gamma_{\overline{D}_l}(\Pi_k) = \lambda_{\overline{D}_l}(\Pi_k) \cdot r_{l,k}$. Hence, $\frac{\gamma_{\overline{D}_l}(\Pi_{k-1})}{\gamma_{\overline{D}_l}(\Pi_k)} = \frac{\lambda_{\overline{D}_l}(\Pi_{k-1})}{\lambda_{\overline{D}_l}(\Pi_k)} \cdot \frac{r_{l,k-1}}{r_{l,k}}$.

From the proof of lemma 1 we know that $\lambda_{\overline{D}_l}(\Pi_k) = \alpha_{l,k-1} \cdot (1 - p_f(\varepsilon, r_{l,k})) \cdot \lambda_{D_l}(\widetilde{r}_{l,k})$,

which implies that: $\frac{\gamma_{\overline{D}_l}(\Pi_{k-1})}{\gamma_{\overline{D}_l}(\Pi_k)} = \frac{\lambda_{D_l}(\widetilde{r}_{l,k-1})}{\lambda_{D_l}(\widetilde{r}_{l,k})} \cdot \frac{r_{l,k-1}}{r_{l,k}} \cdot \frac{1}{1 - p_f(\varepsilon, r_{l,k})}$. Also, the slope

$\lambda_{D_l}(\widetilde{r}_{l,k})$ can be approximated by the derivative of $D_l(\widetilde{r}_{l,k})$. If $D_l(\widetilde{r}_{l,k}) = \varepsilon \sigma^2 2^{-2\widetilde{r}_{l,k}}$, then its derivative is given by: $D_l'(\widetilde{r}_{l,k}) = -2\varepsilon \sigma^2 2^{-2\widetilde{r}_{l,k}} \ln(2)$, implying that:

$$\frac{\gamma_{\overline{D}_l}(\Pi_{k-1})}{\gamma_{\overline{D}_l}(\Pi_k)} = 2^{2(\widetilde{r}_{l,k} - \widetilde{r}_{l,k-1})} \cdot \frac{r_{l,k-1}}{r_{l,k}} \cdot \frac{1}{(1 - p_f(\varepsilon, r_{l,k}))} = 2^{2r_{l,k}} \cdot \frac{r_{l,k-1}}{r_{l,k}} \cdot \frac{1}{1 - p_f(\varepsilon, r_{l,k})} \tag{11}$$

From (11), a sufficient condition such that $\gamma_{\overline{D}_l}(\Pi_k)$ is monotonically decreasing with

k is $2^{2r_{l,k}} \frac{r_{l,k-1}}{r_{l,k}} \geq 1$. Let $r_{l,k} = \beta r_{l,k-1}$. This implies:

$$4^{r_{l,k}} > \frac{r_{l,k}}{r_{l,k-1}} = \beta \Leftrightarrow 4^{\beta r_{l,k-1}} > \beta \Leftrightarrow f(\beta) = \frac{\log_4 \beta}{\beta} < r_{l,k-1} \tag{12}$$

A simple derivation shows that the maximum of $f(\beta)$ is achieved when $\beta = e$. This means that $r_{i,k} > \log_4 e/e = 0.2654$ which ends the proof. Q.E.D.

Based on the above lemmas we are now able to develop our JSCC rate-allocation mechanism. As an example we sketch three possible (convex) paths of the subband distortion evolution when different combinations of code rates are assigned to four packets (see Fig. 1). This example illustrates that one cannot assume the existence a single path out of all the convex paths which delivers the best protection for any number of transmitted packets. We conclude that depending on the number of packets we transmit for each subband, different code rate combinations (different paths) should be considered to find the minimal expected average subband distortion.

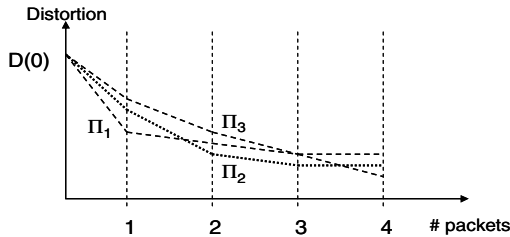


Fig. 1. Example of the subband-distortion paths which can be achieved when subsequent packets are transmitted with different protection levels

With this knowledge, we propose to solve the global optimization problem given by (2) and (3) as follows. For each subband in the wavelet-decomposition we retain from the constructed convex hulls, the paths which result in a minimal distortion at each subsequent packet. Thereby we construct a “virtual” envelope of the convex hulls (see Fig. 2) where each point on the virtual envelope can be achieved through a real path. Based on the two lemmas, it can be proven using simple geometry that this virtual envelope is convex, with monotonically decreasing slopes. Hence, we can perform an optimal rate-allocation in between the subbands by means of classical Lagrangian optimization applied on the virtual hulls computed for each subband.

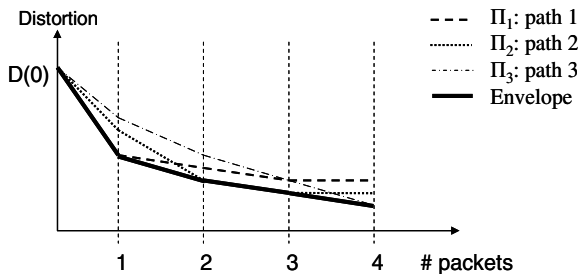


Fig. 2. Example of the derivation of the virtual convex hull of a subband from the real convex hulls of that subband

3.3 Construction of the Convex Hulls: A Computational Analysis

3.3.1 Exhaustive Search with UEP

One way to construct all possible convex hulls under the constraint of UEP is to generate for each subband all possible code rate combinations. From Fig. 3 it can be noticed that this results in an explosive growth of the search space.

The number $y_m(d)$ of different paths to follow in order to achieve protection d at packet m can easily be derived. Let d_{tot} be the number of different code rates and m the number of packets transmitted. Each packet can be protected with a different code and we make the convention that $d = 1$ and $d = d_{tot}$ correspond to the strongest and weakest protection codes respectively.

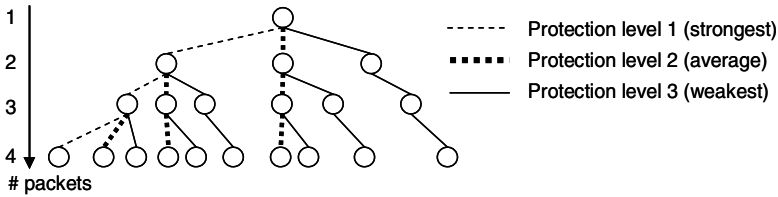


Fig. 3. Example of the explosive growth of the search space when using exhaustive search

It can be shown that:

$$y_m(d) = \sum_{i=1}^d y_{m-1}(i) \tag{13}$$

$$y_m(d) = \frac{d(d+1)(d+2)\dots(d+m-2)}{(m-1)!} = C_{d+m-2}^{d-1} = \binom{d+m-2}{d-1} \tag{14}$$

Proof:

In order to facilitate the proof, we provide an alternate view for all possible paths that are allowed under UEP in Fig. 4. The first packet can be assigned any code rate: $y_1(d) = 1$. The number of paths to follow when 2 packets are transmitted is dependent on the protection of packet 1 and is given by: $y_2(d) = d$, with $1 \leq d \leq d_{tot}$.

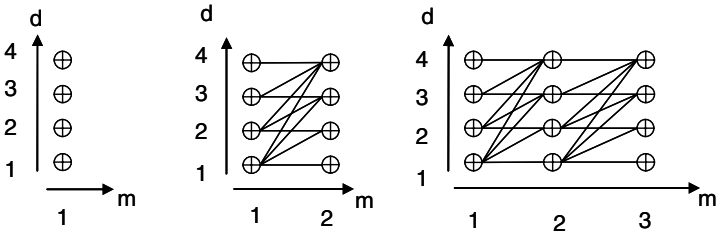


Fig. 4. Example of exhaustive search method with UEP with 4 protection levels and up to 3 transmitted packets

For packet 3, the number of paths depends on the protection of packet 2:

$$y_3(d) = \sum_{i=1}^d y_2(i) = \sum_{i=1}^d i = \frac{d(d+1)}{2},$$

which corresponds to (14) for $m = 3$. Recursively, $y_m(d) = \sum_{i=1}^d y_{m-1}(i)$. Assume by induction that $y_m(d) = \binom{d+m-2}{d-1}$ and let us prove that $y_{m+1}(d) = \binom{d+m-1}{d-1}$. One writes: $y_{m+1}(d) = \sum_{i=1}^d y_m(i) = \sum_{i=1}^d \binom{i+m-2}{i-1}$, which is the form of a known combinatorial sum [19]: $\sum_{i=0}^k \binom{i+p}{i} = \binom{p+k+1}{k}$. With $i' = i - 1$:

$$y_{m+1}(d) = \sum_{i=1}^d y_m(i) = \sum_{i=1}^d \binom{i+m-2}{i-1} = \sum_{i'=0}^{d-1} \binom{i'+m-1}{i'} = \binom{m-1+d-1+1}{d-1} = \binom{m+d-1}{d-1}, \tag{15}$$

which ends the proof.

Q.E.D.

Practically, the above means that for the transmission of, for example, 20 packets with 8 protection levels, 888030 paths should be computed to find the minimal path for one subband. This is computationally much too intensive.

3.3.2 Proposed Simplified Viterbi Search with UEP

In order to significantly reduce the computational burden involved by an exhaustive search approach as described in Section 3.3.1, we propose to continuously eliminate paths that lead to the same protection level for the packet under consideration and let the single path that results in the minimal expected distortion survive. We illustrate our approach in Fig 5. The average expected distortion for each protection path is computed with our recursive formula (10).

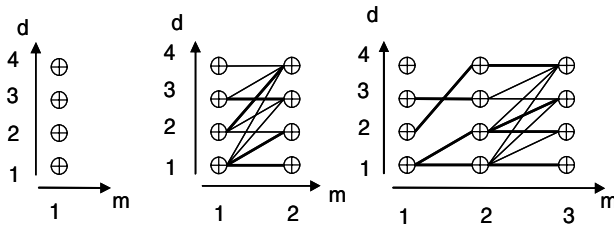


Fig. 5. Proposed approach with limited complexity for the construction of the rate-distortion hulls of a subband. Thin lines represent the computed paths under UEP. Bold lines represent the survivor paths that lead to the minimal distortion for a certain protection level and number of packets transmitted.

The number of additional paths Y_m that need to be computed for the protection of each extra packet with d_{tot} possible protection levels is then given by:

$$Y_1 = d_{tot}, \quad \dots, \quad Y_m = \frac{d_{tot} \cdot (d_{tot} + 1)}{2}$$

The total complexity in terms of the total number of paths to compute when transmitting M_l packets is therefore: $Y = \sum_{m=1}^{M_l} Y_m = d_{tot} + (M_l - 1) \cdot d_{tot} \cdot (d_{tot} + 1) / 2$, which is only of order $O(d^2 \cdot M_l)$. Compared with the algorithm proposed by Banister [14] which presents a complexity of the order of $O(d \cdot M_l^2)$, our algorithm provides a significant reduction in complexity since M_l is typically much larger than d .

4 Coding System Description

4.1 Source Coder

Our source coder consists of a global wavelet transform followed by successive approximation quantization and non-adaptive arithmetic coding. In order to combat the dramatic effect of synchronization loss of the arithmetic coder when an error occurs, we code each subband separately and reset the arithmetic coding process after each bitplane. This results in a small reduction in compression performance but increases the robustness to transmission errors effectively.

The coding process basically consists of a succession of significance and refinement passes [20] performed at each bit-plane. Similar to JPEG-2000, the rate-distortion characteristics of each subband are computed while encoding. With this respect, it can be shown [20] that the reduction in the total distortion, resulting from decoding a wavelet coefficient during the significance pass at bitplane k is given by: $\Delta D_S(k) = 27 \cdot 2^{2k} / 12$. The distortion reduction for the refinement pass of bit-plane k is calculated as: $\Delta D_R = 2^{2(k-1)}$. Finally, the rates spent at each bit-plane and coding pass are the actual rates produced by arithmetic coder.

4.2 Channel Coder

For the protection of the source packets of size $K \times 1$ we use regular LDPC-codes over GF(2) [21]. An LDPC code is a linear block code defined by a sparse parity-check matrix H of dimension $N \times M$ with $M = N - K$. The parity-check matrix H of a *regular* LDPC-code contains a low and fixed number of ones in the rows (also called left or variable degree d_v) and a low and fixed number of ones in the columns (also called right or check degree d_c). In order to create regular (d_v, d_c) LDPC-codes which exhibit good performance, we implemented the Progressive Edge Growth (PEG) algorithm of [22]. This algorithm presents good properties in terms of girth and minimum distance which are dominant criteria to achieve performant codes.

Encoding of the source packets is performed with the dual of the parity-check matrix, called the generator matrix G . We construct the matrix G in systematic form, meaning that the output after encoding includes the original source data. Iterative decoding is performed using a log-domain sum-product algorithm [21]. Since we focus here on fixed-length packets we construct for each different protection level new parity-check and generator matrices with different dimensions and puncture the last redundant bits such that packets of fixed-length n are achieved (with $n < N$).

5 Experimental Results

In our experiments we consider the transmission of embedded image codes in packets of exactly 256 bytes over BECs with 5%, 10%, 20% and 30% of bit erasures. For the channel coding we employ punctured regular (3,6)-LDPC codes of which we measured the performance off-line (see Table 1). The iterative decoding allows up to 100 iterations.

Table 1. Average probability of packet loss for the punctured regular (3,6) LDPC-codes. N is the total number of bytes of the codeword. K is the number of source bytes in the codeword and P is the number of redundant bytes punctured from the codeword.

BEC with 5% losses				BEC with 10% losses				BEC with 20% losses				BEC with 30% losses			
N	K	P	Probability of Packet loss	N	K	P	Probability of Packet loss	N	K	P	Probability of Packet loss	N	K	P	Probability of Packet loss
256	128	0	0.00E+00	256	128	0	0.00E+00	256	128	0	0.00E+00	256	128	0	0.00E+00
414	207	158	0.00E+00	388	194	132	0.00E+00	334	167	78	0.00E+00	284	142	28	0.00E+00
418	209	162	1.00E-06	396	198	140	4.84E-05	342	171	86	4.57E-05	292	146	36	3.12E-05
422	211	166	2.83E-04	400	200	144	1.21E-03	346	173	90	4.12E-04	296	148	40	6.19E-04
426	213	170	3.24E-02	404	202	148	3.63E-02	350	175	94	7.30E-03	300	150	44	6.33E-03
430	215	174	1.84E-01	408	204	152	1.49E-01	356	178	100	1.22E-01	304	152	48	4.04E-02

Table 2. Average performance comparison of (1) exhaustive search, our proposed Viterbi-search with (2) three and (3) all available protection levels for the 512x512 grayscale-images Lena (*left*) and Goldhill (*right*)

		EXHAUSTIVE, UEP SEARCH & 3 PROTECTION LEVELS			VITERBI, UEP SEARCH & 3 PROTECTION LEVELS			VITERBI, UEP SEARCH ALL PROTECTION LEVELS			EXHAUSTIVE, UEP SEARCH & 3 PROTECTION LEVELS			VITERBI, UEP SEARCH & 3 PROTECTION LEVELS			VITERBI, UEP SEARCH ALL PROTECTION LEVELS		
		BEC 5%									BEC 5%								
R_{target} (bpp)	R_{real} (bpp)	R_s (bpp)	PSNR (dB)	R_{real} (bpp)	R_s (bpp)	PSNR (dB)	R_{real} (bpp)	R_s (bpp)	PSNR (dB)	R_{real} (bpp)	R_s (bpp)	PSNR (dB)	R_{real} (bpp)	R_s (bpp)	PSNR (dB)	R_{real} (bpp)	R_s (bpp)	PSNR (dB)	
0.25	0.25	0.205	31.16	0.25	0.205	31.16	0.25	0.205	31.16	0.26	0.211	30.16	0.26	0.211	30.15	0.26	0.211	30.15	
0.5	0.50	0.410	34.38	0.50	0.409	34.37	0.50	0.409	34.37	0.50	0.410	32.68	0.51	0.415	32.73	0.51	0.415	32.73	
1	1.02	0.833	37.80	1.02	0.833	37.80	1.02	0.834	37.81	1.01	0.826	36.28	1.01	0.825	36.27	0.99	0.812	36.18	
		BEC 10%									BEC 10%								
0.25	0.25	0.193	30.85	0.25	0.193	30.85	0.25	0.193	30.85	0.25	0.193	29.89	0.25	0.193	29.89	0.25	0.193	29.89	
0.5	0.50	0.386	34.10	0.51	0.392	34.16	0.51	0.392	34.16	0.50	0.386	32.33	0.50	0.386	32.33	0.50	0.386	32.33	
1	1.07	0.827	37.79	1.07	0.826	37.78	1.08	0.828	37.78	1.02	0.790	36.04	1.02	0.790	36.04	0.99	0.761	35.77	
		BEC 20%									BEC 20%								
0.25	0.25	0.167	30.36	0.25	0.167	30.36	0.25	0.168	30.36	0.25	0.167	29.44	0.25	0.167	29.44	0.25	0.167	29.44	
0.5	0.50	0.335	33.54	0.50	0.334	33.53	0.50	0.334	33.54	0.52	0.346	32.03	0.52	0.346	32.03	0.52	0.346	32.03	
1	1.00	0.668	36.83	1.01	0.670	36.85	1.01	0.670	36.85	1.03	0.690	35.29	1.04	0.692	35.29	1.04	0.693	35.29	
		BEC 30%									BEC 30%								
0.25	0.25	0.143	29.71	0.25	0.143	29.71	0.25	0.143	29.69	0.25	0.143	28.87	0.25	0.143	28.87	0.25	0.143	28.87	
0.5	0.50	0.286	33.14	0.50	0.287	33.13	0.50	0.287	33.14	0.50	0.286	31.53	0.50	0.285	31.53	0.50	0.286	31.53	
1	1.00	0.571	36.17	1.02	0.581	36.23	1.02	0.581	36.23	0.99	0.567	34.62	0.99	0.567	34.62	0.99	0.567	34.63	

Lena 512x512

Goldhill 512x512

In contrast to other JSCC coding approaches [13], [14], [15], we do not include extra CRC-bits to detect errors. This is because the LDPC-codes are linear block codes which present a probability of error detection of practically 100% and which decode the received codewords with a simple matrix multiplication. All our experiments confirm this, since the LDPC-decoder never made false error detection decisions, as this would have led to a de-synchronization of the arithmetic decoder.

We use a five-level (9,7) wavelet-transform and apply embedded coding as described in Section 4.1. The embedded source coding is performed only once and the rate-distortion characteristics of the embedded subband sources are stored as look-up tables. We do not consider the (small) amount of rate used by the header-information in our algorithm and assume that this information reaches the decoder intact.

After source coding, the proposed JSCC-approach determines for given channel characteristics how the subbands should be packetized, i.e. how many packets of each subband are necessary and how each packet should be protected.

Table 2 presents the simulation results when applying (1) an exhaustive search with UEP with three protection levels, (2) our proposed algorithm with three protection levels and (3) our proposed algorithm with all available protection levels on the classical “Lena” and “Goldhill” images (size 512x512 pixels, 8bpp).

The three chosen LDPC-codes are always the three first punctured codes from Table 1, i.e. those for which the number of punctured bytes P is different from zero. In order to have comparable results, the transmission (insertion of random bit erasures) is performed identically for each search method. We repeated the transmission of the packets 5000 times and averaged the resulting PSNRs. This means that equivalent packets in the different search methods are corrupted identically. These results show that our proposed solution yields very near-to-optimal compression performance.

6 Conclusion

In this paper we have introduced a JSCC-system with UEP for the transmission of embedded wavelet-based image codes over binary erasure channels. Our proposed methodology relies on novel proofs concerning the convexity of the distortion evolution in separately encoded subbands. Globally, our proposed approach consists of a simplified Viterbi-search method to define the rate-distortion characteristics of the subbands followed by a global Lagrangian optimization over all subbands. We have showed that our system results in significant complexity reductions while at the same time providing very near-to-optimal compression performance.

Acknowledgments

This research was funded by a PhD grant of the Institute for the Promotion of Innovation through Science and Technology in Flanders (PhD bursary M. R. Stoufs), BELSPO (IAP Phase V—Mobile Multimedia) and the Fund for Scientific Research—Flanders (FWO) (post-doctoral fellowships A. Munteanu and P. Schelkens).

References

1. J. M. Shapiro: Embedded Image Coding Using Zerotrees of Wavelet Coefficients. *Transactions on Signal Processing*, Vol. 41 no. 12 (1993) 3445-3462
2. A. Said, A. W. Pearlman: A New Fast and Efficient Image Codec Based on Set Partitioning in Hierarchical Trees. *IEEE Transactions on Circuits and Systems for Video Technology*, Vol. 6 (1996) 243-250

3. A. Munteanu, J. Cornelis, G. Van der Auwera, P. Cristea: Wavelet Image Compression - the Quadtree Coding Approach. Vol. 3 no. 3 (1999) 176-185
4. C. Berrou, A. Glavieux, P. Thitimajshima: Near Shannon limit error-correcting coding and decoding: Turbo codes. Proceedings of IEEE International Conference on Communications, Geneva, Switzerland, Vol. 2, (1993) 1064-1070
5. R. Gallager: Low-Density Parity-Check Codes. Massachusetts Institute of Technology, PhD thesis, (1963)
6. D. MacKay, R. Neal: Good Codes based on Very Sparse Matrices. Lecture Notes in Computer Science (1995) 100-111
7. H. Jin, A. Khandekar, R. McEliece: Irregular Repeat Accumulate Codes. Proceedings of 2nd International Symposium on Turbo codes and Related Topics, Brest, France, (2000) 1-8
8. M. Luby: LT- codes. Proceedings of 43rd Annual IEEE Symposium on the Foundations of Computer Science (STOC), (2002) 271-280
9. A. Shokrollahi: Raptor codes. Proceedings of IEEE International Symposium on Information Theory, (2003)
10. C. E. Shannon: A mathematical theory of communication. Bell System Technical Journal, Vol. 27 (1948) 379-423
11. J. L. Massey: Joint Source and Channel Coding in Communication Systems and Random Process Theory, Sijthoff & Noordhoff ed. Alphen aan den Rijn, The Netherlands, (1978) 279-293.
12. A. Albanese, J. Blömer, J. Edmonds, M. Luby, M. Sudhan: Priority Encoding Transmission. IEEE Transactions on Information Theory, Vol. 42 no. 6 (1996) 1737-1744
13. Z. Wu, R. Jandhyala, B. Ali, M. W. Marcellin: Joint Source/Channel Coding for Multiple Video Sequences With JPEG2000. IEEE Transactions on Image Processing, Vol. 14 no. 8 (2005) 1020-1032
14. B. A. Banister, B. Belzer, T. R. Fischer: Robust image transmission using JPEG2000 and turbo-codes. IEEE Signal Processing Letters, Vol. 9 (2002) 117-119
15. R. Hamzaoui, V. Stankovic, Z. Xiong: Fast algorithm for distortion-based error protection of embedded image codes. IEEE Transactions on Image Processing, Vol. 14 (2005) 1417-1421
16. V. Stankovic, R. Hamzaoui, D. Saupe: Fast algorithm for rate-based optimal error protection of embedded codes. IEEE Transactions on Communications, Vol. 51 (2003) 1788-1795
17. J. Thie, D. S. Taubman: Optimal Protection Assignment for Scalable Compressed Images. Proceedings of ICIP, (2002) 713-716
18. R. Hamzaoui, V. Stankovic, Z. Xiong: Optimized Error Protection of Scalable Image Bit Streams. IEEE Signal Processing Magazine, Vol. 22 (2005) 91-107
19. D. Zwillinger: CRC, Standard Mathematical Tables and Formulae. 30th edn., Boca Raton (1996)
20. A. Munteanu: Wavelet Image Coding and Multiscale Edge Detection. Department of Electronics and Information Processing, Vrije Universiteit Brussel, Brussel, PhD Thesis, (2003)
21. S. Lin, D. Costello: Error Control Coding: Fundamentals and Applications. 2nd edn. Prentice-Hall, New Jersey (2004)
22. X.-Y. Hu, E. Eleftheriou, D.-M. Arnold: Regular and irregular progressive edge-growth Tanner graphs. IEEE Transactions on Information Theory, Vol. 51 no. 1 (2005) 386-398

Robust Analysis of Silhouettes by Morphological Size Distributions

Olivier Barnich*, Sébastien Jodogne, and Marc Van Droogenbroeck**

University of Liège, Department of Electricity, Electronics and Computer Science,
Institut Montefiore B-28, Sart Tilman, B-4000 Liège, Belgium

Abstract. We address the topic of real-time analysis and recognition of silhouettes. The method that we propose first produces object features obtained by a new type of morphological operators, which can be seen as an extension of existing granulometric filters, and then insert them into a tailored classification scheme.

Intuitively, given a binary segmented image, our operator produces the set of all the largest rectangles that can be wedged inside any connected component of the image. The latter are obtained by a standard background subtraction technique and morphological filtering. To classify connected components into one of the known object categories, the rectangles of a connected component are submitted to a machine learning algorithm called EXtremely RAndomized trees (Extra-trees). The machine learning algorithm is fed with a static database of silhouettes that contains both positive and negative instances. The whole process, including image processing and rectangle classification, is carried out in real-time.

Finally we evaluate our approach on one of today's hot topics: the detection of human silhouettes. We discuss experimental results and show that our method is stable and computationally effective. Therefore, we assess that algorithms like ours introduce new ways for the detection of humans in video sequences.

1 Introduction

During the recent years, the rising of cheap sensors has made of video surveillance a topic of very active research and wide economical interest. In this field, one of the expected major breakthrough would be to design automatic image processing systems able to detect, to track, and to analyze human activities. Unfortunately the amount of data generated by cameras is prohibitively huge, although the informative part of such signals is very tight with respect to their raw content.

Several algorithms in computer vision have been developed to summarize such informative patterns as a set of *visual features*. These algorithms generally rely on the detection of discontinuities in the signal selected by *interest point*

* Olivier Barnich has a grant funded by the FRIA, Belgium.

** This work was supported by the Belgian Walloon Region (<http://www.wallonie.be>), under the CINEMA project.

detectors [1]. Then, a local description of the neighborhood of the interest points is computed [2] and this description serves to track a feature in successive frames of a video sequence. Methods like this, referred to as *local-appearance methods*, have been used with some success in computer vision applications such as image matching, image retrieval, and object recognition (see [3,4]).

From current literature, it is still unclear whether such local-appearance descriptors are appropriate for tracking human silhouettes, or more specifically for gait analysis. Indeed, they are rather computationally expensive, and as they are inherently local, it is impossible for them to represent the overall geometry of a silhouette. There are two potential solutions to this problem: (1) introduce higher-level descriptors able to represent the relative spatial arrangements between visual features [5], or (2) take global appearance (such as contours) into consideration instead of local appearance.

Gait analysis techniques based on the global geometry of the objects have been discussed by BOULGOURIS *et al.* [6]. According to them, techniques that employ binary images are believed to be particularly suited for most practical applications since color or texture informations might not be available or appropriate. The contour of a silhouette is probably the most sensible visual feature in this class. A direct use of it is possible, or it can be transformed into a series of Fourier descriptors as common in shape description. Alternatively the width of silhouette, horizontal and vertical projections, and angular representation are other candidates that have been proposed.

In this paper, we propose a novel approach that is at the crossroad between local- and global-appearance techniques. Our approach innovates in that we propose a new family of visual features that rely on a surfacic description of a silhouette. Intuitively, we cover the silhouette by the set of all the largest rectangles that can be wedged inside of it. More precisely, each (local) position in the silhouette is linked to the subset of the largest rectangles that cover it and that are entirely included in the (global) silhouette.

Surfacic descriptors, like the morphological skeleton [7], have already been studied in the scope of shape compression whose goal is to reduce the amount of redundant information. In general, they require large computation times, which makes them less suitable for real-time applications. This contrasts with our features, as it is possible to compute them in real-time, if enough care is taken in the implementation.

This paper describes an attempt to take advantage of such novel features. To illustrate our approach, we focus on the detection of human bodies in a video stream, like in [8,9]. Basically, we apply *machine learning* algorithms on the rectangles of a silhouette to decide, in real-time, whether this silhouette corresponds to that of a learned instance of a human silhouette. This decision is a compulsory step for any gait recognition task, and improvements in this area will impact on the overall performances of algorithms that deal with the automatic analysis of human behavior. Our results show how promising an approach like ours can be.

The paper is organized as follows. We start by describing the architecture of our silhouettes detection and analysis technique in Section 2, which mainly consists in three steps (silhouettes extraction, description, and classification) respectively detailed in Sections 3, 4, and 5. Experimental results, which consist in the application of our method for the detection of human people in video sequences, are discussed in Section 6.

2 Overall Architecture

The overall architecture of our silhouettes detection, analysis and classification system is depicted in Figure 1. It comprises three main modules.

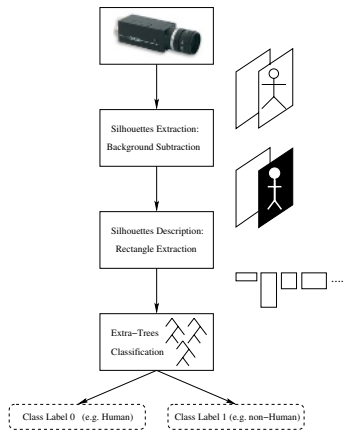


Fig. 1. Overall architecture

1. The first module consists in extracting the candidate silhouettes from the video stream. It is described in Section 3.
2. One of the major difficulties in classification lies in finding appropriate feature measures. In the second module, we use our new granulometric operator to produce a set of features (largest wedged rectangles) describing the extracted silhouettes.
3. The task of the third module is to classify rectangle features to decide whether or not the silhouettes belong to the class of interest. The classification is achieved by the means of an extra-tree learning algorithm as explained in Section 5.

3 Extraction of Silhouettes

The first step of our system consists in the segmentation of the input video stream in order to produce binary silhouettes, which will be fed into the silhouettes

description module. We achieve this by a motion segmentation based on an adaptive background subtraction method.

Background segmentation methods are numerous. The method we have chosen is based on an adaptive modeling of each pixel as a mixture of Gaussians, each of which corresponds to the probability of observing a particular intensity or color for this pixel. In each Gaussian cluster, the mean accounts for the average color or intensity of the pixel, whereas the variance is used to model illumination variations, surface texture, and camera noise. The whole algorithm relies on the assumptions that the background is visible more frequently than the foreground and that its variance is relatively low, which are common assumptions for any background subtraction technique. Extensive description of the algorithm can be found in [10,11] and a tutorial is available at [12]. The technical description is given hereafter.

If X_t is the color or intensity value observed at time t for a particular pixel in the image, the history $\{X_1, \dots, X_t\}$ is modeled as a mixture of K Gaussian distributions. The probability of observing a particular color or intensity value at time t is expressed as

$$P(X_t) = \sum_{i=1}^K \omega_{i,t} \eta(X_t, \mu_{i,t}, \Sigma_{i,t}), \quad (1)$$

where

- K is the number of Gaussian clusters used to model the history of the pixel,
- $\omega_{i,t}$ is the weight associated with the i th cluster at time t –it models the amount of data represented by the i th Gaussian–,
- $\mu_{i,t}$ and $\Sigma_{i,t}$ are the mean and covariance matrix of the i th Gaussian, and
- η is a Gaussian probability density function.

For computational efficiency reasons, the covariance matrix $\Sigma_{i,t}$ is assumed to be isotropic and diagonal

$$\Sigma_{i,t} = \sigma_k^2 \mathbf{I}. \quad (2)$$

The Gaussian distributions are sorted in decreasing order of the ratio $\frac{\omega_{i,t}}{\sigma_{i,t}}$. The j first Gaussians are considered to account for the background, while the rest of them accounts for the foreground. The j factor is dynamically estimated by accumulating the $\omega_{i,t}$ values, according to the computed order of the Gaussians, until a given threshold value T is reached. For this to work we assume that the background is visible more often than the foreground and that its variance is relatively low.

Every new pixel value is checked against the K distributions until a match is found, in which case the pixel receives its class label (background or foreground) according to that of the matched distribution. A match is defined as a pixel value within 2.5 times the standard deviation of a distribution. If no match is found, the pixel is considered as belonging to the foreground. In this case, a new distribution, centered on the pixel color or intensity, is initialized to replace the



Fig. 2. Examples of extracted silhouettes with the Gaussian mixture model background subtraction technique

weakest distribution present in the mixture model. This new distribution is of high initial variance and low prior weight.

Once the new pixel value is classified, the model has to be updated. A standard method would be to use the *expectation maximisation* algorithm. Unfortunately, that would be prohibitively computationally expensive. In [10,11], STAUFFER and GRIMSON give an on-line K -means approximation efficient enough to be performed in real-time on a standard VGA image (640×480 pixels).

After the foreground has been computed, foreground pixels are aggregated by a 8-connected component algorithm. This guarantees that a unique label is assigned to each connected region. Then each connected region is considered as a distinctive input for both the silhouettes description and silhouettes classification modules. Examples of extracted candidate silhouettes by the mixture of Gaussians algorithm are shown on Figure 2.

In the silhouettes description module, each candidate silhouette will be handled as if it was the unique region in the image. There are thus as many silhouettes as connected regions for which an algorithm has to decide whether or not it belongs to a known shape pattern.

4 Features Based on a Granulometric Description by Rectangles

Most surfacic descriptors can be described in terms of the theory of mathematical morphology. Therefore we will use this framework to describe our new feature set.

After a brief introduction to some notations, we will present the framework of granulometries that proved to be the starting point of our development. Then we provide a formal description of our new operator.

4.1 Morphological Operators on Sets

Hereafter we briefly recall some definitions and notations used in mathematical morphology that serves as the framework to define our new feature space. Consider a space \mathcal{E} , which is the continuous Euclidean space \mathbb{R}^n or the discrete space \mathbb{Z}^n , where $n \geq 1$ is an integer. Given a set $X \subseteq \mathcal{E}$ and a vector $b \in \mathcal{E}$, the translate X_b is defined by $X_b = \{x + b \mid x \in X\}$.

Let us take two subsets X and B of \mathcal{E} . We define the $X \oplus B$ and $X \ominus B$ respectively as

$$X \oplus B = \bigcup_{b \in B} X_b = \bigcup_{x \in X} B_x = \{x + b \mid x \in X, b \in B\} \quad (3)$$

$$X \ominus B = \bigcap_{b \in B} X_{-b} = \{p \in \mathcal{E} \mid B_p \subseteq X\}. \quad (4)$$

where B is referred to as the *structuring element*.

When X is eroded by B and then dilated by B , one may end up with a smaller set than the original set X . This set, denoted by $X \circ B$, is called the *opening* of X by B and defined by $X \circ B = (X \ominus B) \oplus B$. The geometric interpretation of an opening is that it is the union of all translated versions B included in X , or in mathematical terms, $X \circ B = \{B_p \mid p \in \mathcal{E}, B_p \subseteq X\}$. Note that this geometrical interpretation is valid for a given set of fixed size. We have to enlarge it to encompass the notion of size or family of structuring elements, which leads us to granulometries.

4.2 Granulometries

The concept of granulometry, introduced by MATHERON [13], is based on the following definition.

Let $\Psi = (\psi_\lambda)_{\lambda \geq 0}$ be a family of image transformations depending on a parameter λ . This family constitutes a granulometry if and only if the following properties are satisfied:

$$\forall \lambda \geq 0, \psi_\lambda \text{ is increasing} \quad (5)$$

$$\forall \lambda \geq 0, \psi_\lambda \text{ is anti-extensive} \quad (6)$$

$$\forall \lambda \geq 0, \mu \geq 0, \psi_\mu \psi_\lambda = \psi_\lambda \psi_\mu = \psi_{\max(\lambda, \mu)}. \quad (7)$$

The third property implies that, for every $\lambda \geq 0$, ψ_λ is an idempotent transformation, that is: $\psi_\lambda \psi_\lambda = \psi_\lambda$. As these properties reflect those of an opening, openings fit nicely in this framework as long as we can order the openings with a scalar. For example, assume that $X \circ rB$ is the opening by a ball of radius r . Then $\Psi = (\psi_r)_{r \geq 0} = (X \circ rB)_{r \geq 0}$ is a granulometry. Of particular interest are granulometries generated by openings by scaled versions of a convex structuring element.

Granulometries, and some measures taken of them, have been applied to problems of texture classification [14], image segmentation, and more recently to the analysis of document images [15].

4.3 Granulometric Curves and Features

MARAGOS [14] has described several useful measurements for granulometries defined by a single scale factor: the *size distribution* and the *pattern spectrum*. The size distribution is a curve that gives the probability of a point belonging to an object to remain into that object after openings with respect to a size factor. The pattern spectrum is defined likewise as the derivative of the size distribution. All

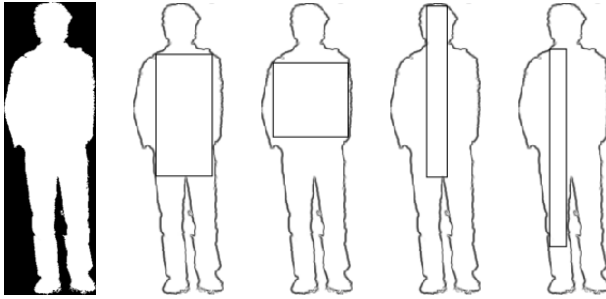


Fig. 3. Examples of largest wedged rectangles contained in a human silhouette

these measures are taken on operator residues driven by a one-dimensional criterion. They are not applicable to a family of arbitrary structuring elements nor are they capable to produce uncorrelated multi-dimensional features. Therefore we define a new operator that produces a *cover*.

Definition 1. [Cover] Let \mathcal{S} be a family of I arbitrary structuring elements $\mathcal{S} = \{S^{i \in I}\}$. A cover of a set X by \mathcal{S} is defined as the union of translated elements of \mathcal{S} that are included in X such that

$$C(X) = \{S_z^j | z \in \mathcal{E} \text{ and } S^j \in \mathcal{S}\} \quad (8)$$

where, if $S_{z'}^{j'}$ and $S_{z''}^{j''}$ both belong to $C(X)$, none of them is totally included in the other one.

As a consequence of this definition, any element of $C(X)$ comprises at least one pixel that uniquely belongs to it. But the upper bound of uniquely covered pixels can be as large as the area of element S^j .

In our application we consider the simplest two-dimensional opening which is of practical interest and practically tractable: an opening by a rectangular structuring element $B = mH \oplus nV$ where mH , nV respectively are m -wide horizontal and n -wide vertical segments. Based on the family \mathcal{B} of all possible rectangle sizes, $C(X)$ will be the union of all the largest non-redundant rectangles included in X . Such rectangles are shown in Figure 3. A fast algorithm for computing this cover is given in [16].

The main advantage of using a cover is that we have a family of structuring elements, describing the surface of an object X , whose members might overlap but all of them uniquely fit somewhere inside of X .

The next step is to extract features from the cover. Since B are rectangles, features like width, height, perimeter, and area spring to mind. For classification purposes however care should be taken to avoid redundant features because it would not increase the performances and could even be counterproductive. For example, HADWIGER [17] has shown that any continuous, additive, and translation and rotation invariant measure on a set X must be a linear combination of the perimeter, area, and EULER-POINCAR number of X .

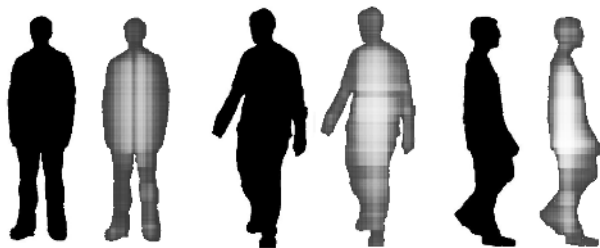


Fig. 4. Examples of rectangles size distributions for human shaped silhouettes. The pixel intensities account for the number of overlapping rectangles that cover each location in the image.

Since translations and scales have some significance in the analysis of silhouettes we reintroduce them by taking the position of elements of $C(X)$ relative to the center of X . Finally we select 5 features on any elements of $C(X)$: width, height¹, 2 relative coordinates of its center, and the percentage of uniquely covered pixels to its area. This last feature is a consequence of taking the cover $C(X)$ to describe X .

5 Silhouettes Classification

Once the set of all the features describing a silhouette has been extracted (see Figure 4 for an illustration of the density of covered pixels), it becomes possible to exploit a machine learning algorithm to map this set into a class. Indeed, such mappings are especially hard to derive by hand and should be *learned* by the system. In our framework, as we are interested in the detection of human silhouettes, only two classes of interest are considered: the class of the human silhouettes, and the class of any other silhouette.

The machine learning approach requires to take two difficulties into account: (1) there is a need of a classifier with excellent generalization abilities not subject to overfitting, and (2) we must define a way to apply this classifier on a set of rectangles, the number of which may widely vary between silhouettes.

To this aim, we propose to use *EXTremely RANdomized trees* (Extra-trees), a fast, yet accurate and versatile machine learning algorithm [18]. Reasons for using extra-trees in our context are threefold: (1) extra-trees have proven successful for solving some color image classification tasks [19], (2) they form a non-parametric function approximation architecture, which do not require previous knowledge, and (3) they have low bias and variance, as well as good performances in generalization.

5.1 Classification Based on Extremely Randomized Trees

We first describe how extra-trees can be used to map a single rectangle to a class. Then we will explain how to map a *set* of rectangles to a class. We will restrict

¹ Note that the perimeter and area derive from the width and height of a rectangle so that it is unnecessary to add them the list of features.

our study of extra-trees to the case where all the input attributes are numerals, which is obviously the case of our rectangular features. Indeed, as mentioned earlier, the input attributes for the rectangles are their width, height, relative positions, and information about the cover.

Intuitively, extra-trees can be thought of as a crossover between *bagging* [20] and *random forests* [21]. They consist of a forest of M independent binary decision trees. Each of their internal nodes is labeled by a threshold on one of the input attributes, that is to be tested in that node. As for the leaves, they are labeled by the classification output. To classify a rectangle through an extra-tree model, this rectangle is independently classified by each tree. This is achieved by starting at the root node, then progressing down the tree according to the result of the tests on the threshold found during the descent, until a leaf is reached. Doing so, each sub-tree votes for a class. Finally, the class that obtains the majority of votes is assigned to the rectangle.

The sub-trees are built in a top-down fashion, by successively splitting the leaf nodes where the output variable does vary. For each input variable, the algorithm computes its variation bounds and uniformly chooses one random threshold between those bounds –this is similar to the case of random forests. Once a threshold has been chosen for every input variable, the split that gives the best information-theoretic score on the classification output is kept –this is similar to bagging. This will guarantee that the variance in the model is reduced (thanks to the presence of a forest of independent sub-trees), as well as bias (thanks to the random selection of the thresholds), while taking advantage of an information measure that guides the search for good splits.

5.2 Classification of Silhouettes

We have just described the process of classifying *one* rectangle. But we describe a silhouette X by its cover $C(X)$ which is a set of rectangles. Furthermore, two distinct silhouettes can have a different number of rectangles inside them. We must therefore introduce a meta-rule over the extra-trees for mapping a set $C(X)$ to a class. In this work, we exploit an idea that is similar to that of MARE *et al.*, which was used in the context of image classification [19].

Let M be a fixed positive integer. Given the set $C(X)$ of rectangles that shapes the silhouette X , we select the first M rectangles inside this set, which induces a subset $C_M(X) \subseteq C(X)$. Then, we apply the extra-trees model onto each rectangle inside $C_M(X)$. This process generates one vote per rectangle. Finally, the silhouette X is mapped into the class that has obtained the majority of the votes.

6 Experimental Results

6.1 Dataset Collection

As mentioned in the introduction, we have focused our experiments on the detection of human silhouettes in a video stream. The extra-trees have been trained

² This line determines the strength of the attribute selection process. The choice of \sqrt{n} is discussed in [18].



Fig. 5. A few examples of negative instances contained in the training dataset



Fig. 6. Subset of positive instances contained in the training dataset

on a dataset of silhouettes that contains both silhouettes of human bodies and silhouettes of other kind of objects. We have fed the learning set with a large number of instances for each of those two classes.

Some instances of non-human silhouettes, called negative instances, are displayed in Figure 5. The negative samples are the union of non-human silhouettes that were extracted from a live video stream by the background subtraction technique presented in Section 3, and of images that were taken from the COIL-100 database [22]. There are about 12,000 images in this dataset. As for the positive instances, we have about 3,000 human silhouettes. Some of them are represented in Figure 6. Those two datasets have been converted to a database that has been fed into the extra-trees learning algorithm (cf. Section 5).

6.2 Tests on Real-World Images

We have tested our algorithms on a color video stream of 640×480 pixels that was captured with a FireWire CCD camera. The whole process (including silhouettes extraction, description, and classification) was carried out at approximately five frames per second on a Pentium IV computer at 3.4 GHz.



Fig. 7. Examples of silhouettes classified correctly. A white frame around an object indicates that the system classifies it as a human silhouette.

The detection of human silhouettes is very robust since the number of correct classifications largely outnumbers misclassifications, although we ignored any correlation between successive frames. Example images of correct (resp. wrong) classifications are shown in Figure 7 (resp. in Figure 8). Our method might be subject to improvements, one of them being the use of a prediction scheme



Fig. 8. Examples of misclassified silhouettes

between successive frames, but these first results demonstrate that on single images our system is capable to recognize specific silhouettes in a semi-controlled environment.

7 Conclusions

In this paper we propose a new system for the real-time detection and classification of binary silhouettes. The silhouettes are extracted from an input video stream using a standard background subtraction algorithm. Then, each silhouette is treated by a new kind of granulometric filter that produces a morphological cover of the silhouette and characterizes it as the set of all the largest rectangles that can be wedged inside of it. One of the major achievements is that we managed to implement the feature extraction step in real-time, which is uncommon for surface-based descriptors. The rectangle features are then fed into an extra-trees classifier that assigns a class label to each detected silhouette. Thanks to the simple tree-based structure of extra-trees, the classification step is also very fast. As a consequence, the whole process that consists of silhouettes detection, analysis and classification can be carried out in real-time on a common computer.

Empirical results that consisted in the application of our method to images captured with a CCD camera put in an environment unknown to the learning process show that our method manages to detect human silhouettes with a high level of confidence. Future work will feature a systematic evaluation of the performances of our approach. We will also investigate its exploitation in more complex tasks such as gait recognition, human tracking, or even general object tracking.

References

1. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. *International Journal of Computer Vision* **37** (2000) 151–172
2. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. In: *Proc. of the IEEE Conference on Computer Vision and Pattern Recognition*. Volume 2., Madison (WI, USA) (2003) 257–263
3. Schmid, C., Mohr, R.: Local greyvalue invariants for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19** (1997) 530–535
4. Lowe, D.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* **60** (2004) 91–110

5. Mathes, T., Piater, J.: Robust non-rigid object tracking using point distribution models. In: Proc. of the British Machine Vision Conference, Oxford (UK) (2005) 849–858
6. Boulgouris, N., Hatzinakos, D., Plataniotis, K.: Gait recognition: a challenging signal processing technology for biometric identification. *ispmag* **22** (2005) 78–90
7. Serra, J.: Image analysis and mathematical morphology. Academic Press, New York (1982)
8. Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., Poggio, T.: Pedestrian detection using wavelet templates. In: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition, Washington, DC, USA, IEEE Computer Society (1997) 193
9. Wren, C.R., Azarbayejani, A., Darrell, T., Pentland, A.: Pfinder: Real-time tracking of the human body. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **19** (1997) 780–785
10. Stauffer, C., Grimson, E.: Adaptive background mixture models for real-time tracking. In: Proc. of the IEEE Conference on Computer Vision and Pattern Recognition. (1999) 246–252
11. Stauffer, C., Grimson, E.: Learning patterns of activity using real-time tracking. *IEEE Trans. Pattern Anal. Mach. Intell.* **22** (2000) 747–757
12. Power, P., Schoonees, J.: Understanding background mixture models for foreground segmentation. In: Proc. Images and Vision Computing, Auckland, NZ (2002)
13. Matheron, G.: *Éléments pour une théorie des milieux poreux*. Masson, Paris (1967)
14. Maragos, P.: Pattern spectrum and multiscale shape representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **11** (1989) 701–716
15. Bagdanov, A., Worring, M.: Granulometric analysis of document images. In IEEE, ed.: Proceedings of the International Conference on Pattern Recognition, Volume I. (2002) 478–481
16. Van Droogenbroeck, M.: Algorithms for openings of binary and label images with rectangular structuring elements. In Talbot, H., Beare, R., eds.: *Mathematical morphology*. CSIRO Publishing, Sydney, Australia (2002) 197–207
17. Hadwiger, H.: *Vorlesungen über inhalt, oberfläche and isoperimetric*. Springer Verlag (1957)
18. Geurts, P., Ernst, D., Wehenkel, L.: Extremely randomized trees. To appear in *Machine Learning Journal* (2006) Available for download at <http://www.montefiore.ulg.ac.be/services/stochastic/pubs/2006/GEW06a/>.
19. Marée, R., Geurts, P., Piater, J., Wehenkel, L.: Random subwindows for robust image classification. In: IEEE Conference on Computer Vision and Pattern Recognition. Volume 1., San Diego (CA, USA) (2005) 34–40
20. Breiman, L.: Bagging predictors. *Machine Learning* **26** (1996) 123–140
21. Breiman, L.: Random forests. *Machine Learning* **45** (2001) 5–32
22. Nene, S., Nayar, S., Murase, H.: Columbia object image library: Coil-100. Technical report (1996)

Enhanced Watermarking Scheme Based on Texture Analysis

Ivan O. Lopes¹, Celia A.Z. Barcelos²,
Marcos A. Batista³, and Anselmo M. Silva¹

¹ Federal University of Uberlândia - FACOM, Faculty of Computer Science
38400-902 Uberlândia MG, Brazil

ivan@pos.facom.ufu.br, anselmo@comp.ufu.br

² Federal University of Uberlândia - FAMAT, Faculty of Mathematics
38400-902 Uberlândia MG, Brazil

celiazb@ufu.br

³ Federal University of Goiás - CAC, Department of Computer Science
75700-000 Catalão GO, Brazil

marcos@catalao.ufg.br

Abstract. This paper proposes a new approach in digital watermarking applications that can be adapted for embedding either fragile or robust watermarking in a digital image in the spatial domain or in the frequency domain. The main objective of the proposed scheme is to explore the amount of texture or edge pixels belonging to the host image in order to insert more information while preserving the robustness of the scheme without degrading the visual quality of the watermarked image. The host image is divided into blocks and each block can be subdivided into sub-blocks according to its texture analysis. The number of sub-blocks that each block will be divided into depends on the amount of texture or edge pixels presented by it. The numerical results show that the proposed scheme is better in JPEG compression attacks, and far exceeds others in watermark size capacity.

Keywords: Watermarking, Texture Analysis, Copyright Protection, Authentication.

1 Introduction

Digital technologies are ever more present in society. The Internet provides a cost effective way of exchanging information, however the security concerned with data authenticity and copyright are not always guaranteed. The digital watermarking techniques emerge as an effective means for protecting intellectual property.

According to Cox et al [4], “watermarking is the practice of changing someone’s work, in an imperceptible way, by introducing a message into it”.

Digital watermarking methods can be separated into two different groups: robust and fragile; and the procedure of watermark encoding can be carried out in the spatial domain [6], [8], or in the frequency domain [5], [2].

In the spatial domain, the watermark embedding into the host image can be done by simply changing its LSB's (Least Significant Bit) when dealing with binary watermarks; these modifications can be easily detected by some kind of image processing operations [7]. These techniques are commonly used for fragile watermark "authentication" [3].

The watermarking embedding into the frequency domain shows, a more robust watermarked image result than that of the usual image processing operations which use fragile ones, and, for this reason, they are generally used for copyright protection. One of the main drawbacks is that the watermark to be inserted has restrictions on its size in order to preserve the visual quality of the watermarked image.

To solve this problem, a method is presented. This method is able to insert binary watermarks containing more information than the original method, and preserves the visual quality of the watermarked image, maintaining its robustness.

A texture image analysis was considered and introduced in the Wu and Shih scheme [9], with the objective of increasing the amount of information to be inserted without the degradation of the watermarked image visual quality. In the proposed technique, watermark is embedded into significant areas according to the amount of information presented by each area.

This paper is organized as follows: Section 2 presents some correlated methods, section 3 presents the proposed method, and the obtained results are presented in section 4. The conclusion is presented in section 5.

2 Correlated Work

The Wu and Shih's scheme [9] uses as its inspiration two well known methods in the watermark area, when dealing with binary watermark. The Wong's method [8] has the objective of inserting watermarks for authentication, and the Cox's method [2] uses the watermark to protect copyrights.

The Wong scheme [8] presents a block-based fragile watermarking technique by adopting the RSA (Rivest-Shamir-Adleman) public key encryption algorithm and Message-Digest-5 for the hashing function.

The Cox technique [2], proposed embedding a watermark into the frequency domain of the host image, inserting the watermarking into the highest coefficients of the host image.

The method proposed by Wu and Shih [9] has as its main objective the capacity to embed a fragile watermark or robust watermark in a unique system.

The authors used two parameters, a matrix QF (Quantify Factor) and $VSTW$ (Varying Sized Transform Window) to reach their integration purpose. The matrix QF determines if the process will be fragile or robust, and the $VSTW$ parameter determines if the process will be performed in the spatial or frequency domain.

Following the ideas presented in Wu and Shih's method [9], and trying to increase the quantity of information embedded in the host image while preserving

the visual quality of the watermarked image and the robustness a new technique is proposed.

3 The Proposed Watermarking Scheme

The proposed method is an improvement to the Wu and Shih's method, which takes into account the texture analysis in order to sub-divide the initial block image division. For the understanding of the proposed scheme some details will follow.

In Wu and Shih's method [9], the watermark is embedded at the most significant point of each block.

The insertion into the most significant points makes the watermark size limited to the number of blocks that the host image was divided.

If we have a host image H of size 256×256 pixels and divide it into blocks of size 16×16 , the information to be embedded will be of size at most 256 pixels. If the host image H is divided in smaller blocks, the watermarked image visual quality can deteriorate. We can solve this problem; by lowering some values of the matrix QF but this procedure will diminish the robustness of the method.

Each block of the host image has one QF which is a matrix of the same size as the block size. Each position of the matrix QF informs the pixel what will be changed. The value belonging to this position indicates the bit which will store the information. The possible values for the QF position are $2k, k \in \mathbb{N}$.

The host image is divided into blocks and each block is subdivided taking into consideration its entropy and/or edge pixels. As entropy (or edge pixels) increase the number of sub-block can be increased, allowing an increase in the amount of information to be inserted.

The encoding and decoding procedure of the proposed method are described below.

3.1 Embedding Algorithm

Let H be a gray-level host image, and let W be a binary watermark image of size $M \times P$. Let H_B be the block-based image obtained by splitting H into non-overlapping blocks. The entropy (or the edge point's quantity) will be used for the block division decision and for the amount of information that each sub-block will receive. Let HB_k the k^{th} sub-block and QF_k be the matrix correspondent quantify factor.

As in the Wu and Shih's scheme, in order to increase the security of the process, a composition of the watermark with the pixel based (PB) features of H will be used which will be inserted into the host image H . The PB is a matrix pixel-based feature extracted from H using morphological operations, and is used the moment that the original image and the watermarked image have almost the same PB , in general [9].

The algorithm proceeds as follows:

Step 1 - Obtaining PB from H :

Create a structure element S as in [9];

Obtain H^D , as a dilation of H by S ;

Obtain H^E , as an erode of H by S ;

For each (i, j) ,

If

$$\left(0 \leq \frac{H(i, j) - H^E(i, j)}{H^D(i, j) - H^E(i, j)} \leq T_1 \right)$$

or

$$\left(T_2 \leq \frac{H(i, j) - H^E(i, j)}{H^D(i, j) - H^E(i, j)} \leq 1 \right)$$

then $PB(i, j) = 1$;

else $PB(i, j) = 0$.

Note: The values of T_1 and T_2 used in the experiments were 0.1 and 0.9, respectively.

Step 2 - By joining the original watermark W with the PB in order to obtain FW :

For $i = 1, \dots, M$ and $j = 1, \dots, P$ do

$$FW(i, j) = PB(i, j) \text{ XOR } W(i, j).$$

Step 3 - Obtaining the blocks H_B :

Split H into non-overlapping blocks H_B .

Step 4 - Defining HB set:

An edge detector and/or the image entropy calculus of each block H_B is performed and, if the result of edge/texture level is greater than a given threshold T , this block is sub-divided defining at the end of this step, the HB set.

Step 5 - Obtaining HB_k^{DCT} :

Compute the DCT of each block HB_k .

Step 6 - Obtaining HB_k^Q :

In each HB_k execute:

For each position (i, j) of H calculated

$$HB_k^Q(i, j) = \frac{HB_k^{DCT}(i, j)}{QF_k(i, j)}.$$

Step 7 - Insertion of FW :

Let $(m, p) = (1, 1)$

For each k execute:

For each position (i, j) of HB_k , calculate

$$HB_k^{WF}(i, j) = HB_k^Q(i, j)$$

If $QF_k(i, j) > 1$ do:

$LSB \{HB_k^{WF}(i, j)\} = FW(m, p)$
 $p = p + 1$
 If $p > P$, do
 $p = 1$
 $m = m + 1$.

Step 8 - Obtaining HB^{WMF} :

For each HB_k do

For each position (i, j) of H calculate

$$HB_k^{WMF}(i, j) = HB_k^{WF}(i, j) \cdot QF_k(i, j).$$

Step 9 - Obtaining HB^{WS} :

Compute the $IDCT$ of each HB_k^{WMF} .

Step 10 - Obtaining the watermarked image HW :

HW will be formed with all the sub-blocks HB_k^{WS} distributed in the same position as the correspondent block HB_k as occupied in H .

In the extraction process, some input variables are used. Which are: The watermarked image HW , the same S structure element, the same QF matrix, and the same threshold T_1 and T_2 used in the insertion process.

3.2 Decoding Algorithm

Step 1 - Obtaining PB from HW :

Use the same structure element S from the insertion process;

Obtain HW^D , as a dilation of HW by S ;

Obtain HW^E , as an erode of HW by S ;

For each (i, j) ,

If

$$\left(0 \leq \frac{HW(i, j) - HW^E(i, j)}{HW^D(i, j) - HW^E(i, j)} \leq T_1 \right)$$

or

$$\left(T_2 \leq \frac{HW(i, j) - HW^E(i, j)}{HW^D(i, j) - HW^E(i, j)} \leq 1 \right)$$

then $PB(i, j) = 1$;

else $PB(i, j) = 0$.

Step 2 - Obtaining the blocks HW_B :

Split HB^{WS} into non-overlapping blocks.

Step 3 - Defining HWB set:

An edge detector and/or image entropy calculating of each block HW_B is performed and if the result of edge/texture level is greater than a given threshold T , this block is sub-divided. This procedure define, at the end of this step, the HWB set.

Step 4 - Obtaining HWB_k^{DCT} :
 Compute the DCT of each HWB_k .

Step 5 - Obtaining HWB_k^Q :
 For each HWB_k do
 For each position (i, j) of HWB calculate

$$HWB_k^Q(i, j) = \frac{HWB_k^{DCT}(i, j)}{QF_k(i, j)}.$$

Step 6 - Obtaining HWB^{LSB} :
 Let $(m, p) = (1, 1)$
 For each k do
 For each position (i, j) of HWB_k^Q , do
 If $QF_k(i, j) > 1$ do:

$$HWB^{LSB}(m, p) = \text{LSB} \{HWB_k^Q(i, j)\}$$

$$p = p + 1$$
 If $p > P$, do

$$p = 1$$

$$m = m + 1.$$

Step 7 - Compound the image HWB^{LSB} with the PB in order to obtain W' :
 For $i = 1, \dots, M$ and $j = 1, \dots, P$

$$W'(i, j) = PB(i, j) \text{ XOR } HWB^{LSB}(i, j).$$

4 Experimental Results

To test the proposed scheme a system for the embedding and detection of a watermark W was implemented. Some tests were (executed) to compare the proposed method with Wu and Shih's method performance.

The value of $QF(1, 1)$ which determines the watermark's robustness, in general receives the value 2^k . The best choice is always the highest k which does not cause image degradation.

The same k value applied to different images, does not necessarily produce the same robustness, or be it, the choice of a k constant, which is effective in the balance of robustness of the technique versus visual degradation, depends entirely upon the image's properties, such as edges, textures, etc.

In the two reported experiments, the highest k possible was chosen which did not visual deteriorate the host image.

In the first experiment Lenna's picture (size 256×256) was split into blocks of size 8×8 . In Wu and Shih's method the watermark W , with size 16×64 , was embedded at the 6^{th} LSB, i.e., $QF(1, 1) = 32$ and $QF(i, j) = 1$ for all $(i, j) \neq (1, 1)$. In the proposed scheme, the host image Lenna was first split into blocks of size 8×8 and, after an entropy image analysis, some blocks were sub-split into blocks of size 4×4 , then the inserted watermark could be larger than that used in Wu's method. The size used was 17×77 , i.e; with 1309 pixels of information, which were embedded at the 6^{th} LSB.

In the second experiment, another test was performed with both methods. A fruit picture (size 256×256) was split into blocks of size 8×8 and the watermark of size 16×64 was embedded at the 5th LSB, i.e., $QF(1, 1) = 16$ and $QF(i, j) = 1$ for all $(i, j) \neq (1, 1)$, in the Wu's technique. Using the same QF in the proposed scheme, the fruit picture (size 256×256) was split into blocks and sub-blocks of size 8×8 and 4×4 , respectively. A watermark of size 20×71 , with 1420 pixels of information, was inserted at the 5th LSB in the host image.

Fig. 1 illustrates the results obtained when the compression JPEG were applied in the robust watermark insertion.

Fig. 1(a) and 1(b) are images of size 256×256 , watermarked by Wu and Shih's method and the proposed method, and which were given a JPEG compression attack of 40%, respectively, figure 1(c) is the original watermark of size 16×64 and figures 1(e), 1(f) and 1(g) are extracted watermarks of the watermarked images by Wu and Shih's method and which suffer JPEG compression of 90, 70 and 40%, respectively, when the Lenna picture was used. Fig. 1(d) is the original watermark of size 17×77 and figures 1(h), 1(i) and 1(j) are extracted watermarks of the watermarked images by the proposed method and which suffer JPEG compression of 90, 70 and 40%, respectively, when the Lenna picture was used. Figure 1(n) is the original watermark of size 16×64 and figures 1(p), 1(q) and 1(r) are extracted watermarks of the watermarked images by Wu and Shih's method and which suffer JPEG compression of 90, 70 and 40% respectively, when the Fruits picture was used. Fig. 1(o) is the original watermark of size 20×71 and figures 1(s), 1(t) and 1(u) are extracted watermarks by of the watermarked images by proposed method and which suffer JPEG compression of 90, 70 and 40% respectively, when the Fruits picture was used.

To measure the similarity between the original watermark W and the extracted ones W' a normalized correlation coefficient was used.

$$NC(W, W') = \frac{\sum_{i=1}^N W_i W'_i}{\sqrt{\sum_{i=1}^N W_i^2} \sqrt{\sum_{i=1}^N W_i'^2}} \quad (1)$$

where $W = (W_1, W_2, \dots, W_N)$ and $W' = (W'_1, W'_2, \dots, W'_N)$.

The results illustrate that the proposed method has almost the same robustness of the Wu and Shih's method, even when the watermark has around 30 to 40% more information pixels than the prior watermark keeping the same visual quality.

Table 1 presents the correlation values obtained using the extracted watermarks obtained by both schemes, and are shown in Fig. 1.

In experiment 1 where the Lenna picture was used, the similarity between the original watermark and the watermark extracted from the image which suffered JPEG compression of 90% it was 0.8346 in the Wu and Shih's method and was 0.8573 in the proposed method. When the watermarked image suffers JPEG compression of 40% the similarity was 0.5880 in the Wu and Shih's method and was 0.6490 in the proposed method.

In experiment 2 where the Fruits picture was used, the similarity between the original watermark and the watermark extracted from the image which suffered

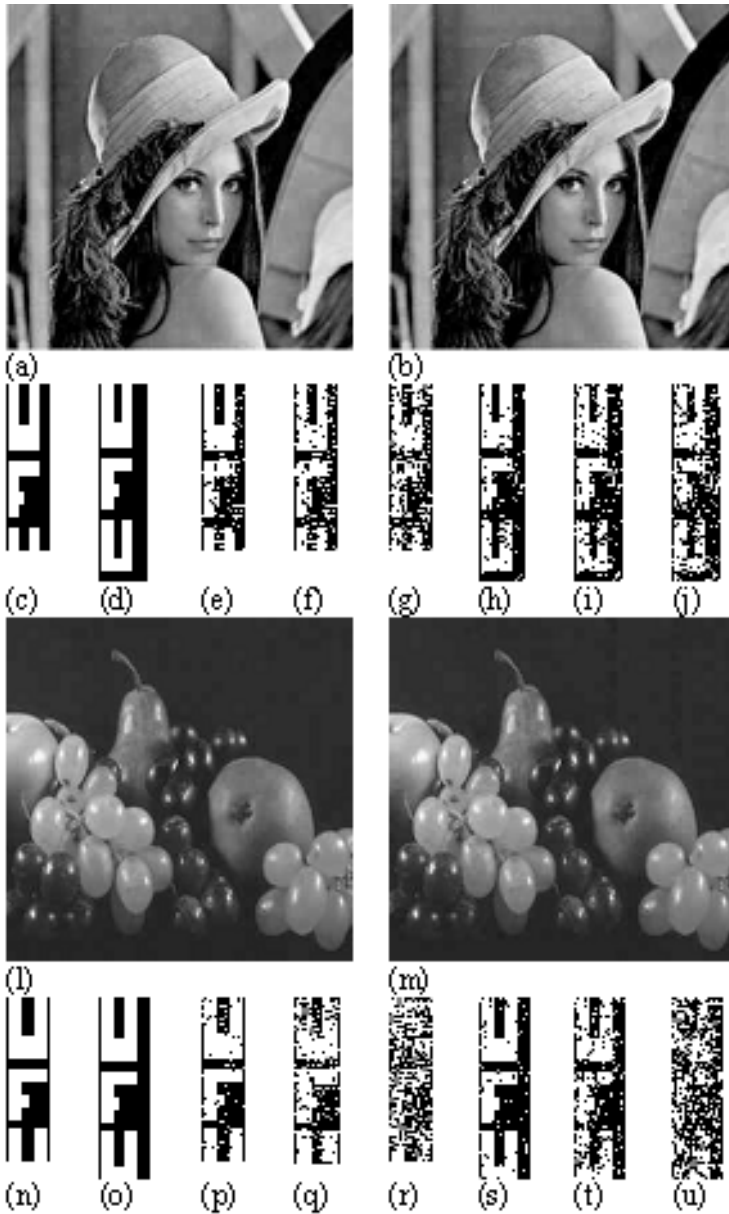


Fig. 1. (a) and (l) are watermarked images by Wu and Shih's method, (c) and (n) are watermark used by Wu and Shih's method, (b) and (m) are watermarked images by proposed method, (d) and (o) are watermark used by proposed method, (e) (f) (g) (p) (q) and (r) are extracted watermarks by Wu and Shih's method, (h) (i) (j) (s) (t) and (u) are extracted watermarks by proposed method

Table 1. The correlation values between extracted and original watermarks

Wu and Shih's Method	Correlation Value
Original Watermark (c) X Extracted Watermark (e)	0.8346
Original Watermark (c) X Extracted Watermark (f)	0.7381
Original Watermark (c) X Extracted Watermark (g)	0.5880
Original Watermark (n) X Extracted Watermark (p)	0.8797
Original Watermark (n) X Extracted Watermark (q)	0.6844
Original Watermark (n) X Extracted Watermark (r)	0.2577
Proposed Method	Correlation Value
Original Watermark (d) X Extracted Watermark (h)	0.8573
Original Watermark (d) X Extracted Watermark (i)	0.7543
Original Watermark (d) X Extracted Watermark (j)	0.6490
Original Watermark (o) X Extracted Watermark (s)	0.8924
Original Watermark (o) X Extracted Watermark (t)	0.7049
Original Watermark (o) X Extracted Watermark (u)	0.3155



Fig. 2. (a) Watermarked image, (b) Watermark, (c) Attacked image and (d) Extracted watermark

JPEG compression of 90% it was 0.8797 in the Wu and Shih's method and was 0.8924 in the proposed method. When the watermarked image suffers JPEG compression of 40% the similarity was 0.2577 in the Wu and Shih's method and was 0.3155 in the proposed method.

The robustness was less in the second experiments as the alterations were realized in the 5th LSB and not in the 6th as in experiment 1, this is because the original image used in the second experiment possess a very large homogeneous area, in this manner if the insertions were to be realized in the 6th LSB, the visual quality of the image would suffer a considerable visual degradation. The insertions in the 5th and 6th LSB determine if the watermark is going to be more or less robust. The greater the robustness of the method, the higher the visual quality degradation of the watermarked image, therefore, the lower the robustness of the method the higher the visual quality of the watermarked image, in this way each image should be analyzed so that desired values between robustness and visual quality can be obtained.

Fig. 2, shows a watermarking application to illustrate that the proposed method can also be used for fragile watermarking purposes. The Lenna's picture, presented in figure 2(a), was divided into blocks of size 1×1 and $QF(i, j)=1$, for all (i, j) . The watermark Federal University of Uberlândia Logotype of size 256×256 , presented in figure 2(b), was inserted. The obtained results are the same in both of the methods once that, there are no differences between them.

5 Conclusion

An enhanced watermarking scheme based on texture analysis is proposed in this paper. The scheme can be seen as a modification of the Wu's method. A texture analysis followed the sub-splitting of each Wu's block which makes the proposed scheme more robust when compared with Wu's scheme once it allows for the insertion of a larger watermark with the same visual quality and same JPEG attack resistance.

When dealing with fragile watermarking both schemes have the same performance. The proposed scheme can be adjusted to fragile or robust watermarking, as desired by the user, choosing the constant $VSTW$ and the matrix QF , this means, taking into account the number of the blocks used to split the host image, and the choice of the position and the intensity of the pixel to be modified in the watermark insertion process.

References

1. Craver, S., Memon, N., Yeo, B., Yeung, M. M.: Resolving Rightful Ownerships with Invisible Watermarking Techniques: Limitations, Attacks, and Implications. *IEEE Journal on Selected Areas in Communications*, Vol. 16. 4 (1998).
2. Cox, I. J., Killian, J., Leighton, F. T., Shamoon, T.: Secure Spread Spectrum Watermarking for Multimedia. *IEEE Trans. Image Process*, Vol. 6 12 (1997) 1673–1687.
3. Cox, I. J., Miller, M. L., Bloom, J. A.: Watermarking application and their properties. Published in the *Int. Conf. On Information Technology*, Las Vegas (2000).

4. Cox, I. J., Miller, M. L., Bloom, J. A.: Digital Watermarking. 2nd edn. Morgan Kaufmann Publishers (2001).
5. Kim, B., Choi, J., Park, C., Won, J., Kwak, D., Oh, S., Koh, C., Park, K.: Robust digital image watermarking method against geometrical attacks. *Real-Time Imaging*, Vol. 9 (2003) 139–149.
6. Kim, H. Y.: Marcas d'gua Frgeis de Autenticao pra Imagens em Tonalidade Contnua e Esteganografia para Imagens Binrias e Meio-Tom. *RITA*, Vol. 8 1 (2001).
7. Swanson, M. D., Zhu, B., Tewfik, A. H.: Transparent Robust Image Watermarking. *SPIE Conference on Visual Communications and Image Processing*, (1996).
8. Wong, P. W.: A Public Key Watermarking for Image Verification and Authentication. in: *Proceeding of the IEEE International Conference on Image Processing*, Chicago, IL (1998) 425–429.
9. Wu, Y. T., Shih, F. Y.: An Adjusted-purpose Digital Watermarking Technique. *Pattern Recognition*, Vol. 37 (2004) 2349–2359.

A Robust Watermarking Algorithm Using Attack Pattern Analysis

Dongeun Lee¹, Taekyung Kim¹, Seongwon Lee^{2,*}, and Joonki Paik¹

¹ Image Processing and Intelligent Systems Laboratory, Department of Image Engineering,
Graduate School of Advanced Imaging Science, Multimedia, and Film,
Chung-Ang University, Seoul, Korea
ehddms98@wm.cau.ac.kr
<http://ipis.cau.ac.kr>

² Department of Computer Engineering, College of Electronics and Information,
Kwangwoon University, Seoul, Korea
swlee@kw.ac.kr

Abstract. In this paper we propose a method that analyzes attack patterns and extracts watermark after restoring the watermarked image from the geometric attacks. The proposed algorithm consists of a spatial-domain key insertion part for attack analysis and a frequency-domain watermark insertion part using discrete wavelet transform. With the spatial-domain key extracted from the damaged image, the proposed algorithm analyzes distortion and finds the attack pattern. After restoring the damaged image, the algorithm extracts the embedded watermark. By using both spatial domain key and frequency domain watermark, the proposed algorithm can achieve robust watermark extraction against geometrical attacks and image compressions such as JPEG.

1 Introduction

Digital watermarking is a digital content copyright protection technique against unauthorized uses of multimedia contents such as illegal copy, distribution, and forgery. Digital watermarking inserts and extracts copyright information called watermark into the digital contents to prove the ownership of the copyright holder. The watermarking techniques slightly modify the original data during the watermark insertion. The watermark can be either visible or invisible. In case of invisible watermarking, the watermarked image should be indistinguishable from the original image. The watermarking are also divided into fragile watermarking to show the existence of illegal modification (often called attack) and robust watermarking that the watermark endures attacks and noise.

Watermark insertion can be done in spatial domain or in frequency domain. The spatial watermark insertion manipulates image pixels, especially on least significant bits that have less perceptual effect on the image. Although the special watermark insertion is simple and easy to implement, it is weak at attacks and noise. On the other hand, the frequency domain watermark insertion that is robust at attacks is performed

* Corresponding author.

to the frequency coefficients of the image. *DFT* (discrete Fourier transform)[1,2], *DCT* (discrete cosine transform), and *DWT* (discrete wavelet transform) are used in the frequency domain watermark insertion[3-5].

Although some significant progresses have been made recently, one of major problems in the practical watermarking technology is the insufficient robustness of existing watermarking algorithms against image compression such as JPEG2000 and h.264 [6, 7] and geometrical distortions such as translation, rotation, scaling, cropping, change of aspect ratio, and shearing. These geometrical distortions cause the loss of geometric synchronization that is necessary in watermark detection and decoding [8]. There are two different types of solutions to resisting geometrical attacks: non blind and blind methods [9]. With the nonblind approach, due to availability of the original image, the problem can be resolved with a good solution by effective search between the geometrically attacked and unattacked image [10, 11]. The blind solution that extracts watermark without the original image has wider applications but is obviously more challenging.

In this paper, we propose a watermarking algorithm that is robust at various attack patterns by using both a spatial domain key and a DWT domain watermark. In addition, the proposed algorithm provides the ability to find attack patterns in the watermarked image. We evaluate the performance of the proposed algorithm in aspects of image quality, robustness, and attack analysis. In Section 2, a brief introduction of wavelet transform and the proposed watermarking algorithm is presented. The performance of the proposed algorithm is evaluated in Section 3. We conclude the paper in Section 4.

2 Proposed Watermarking Method

The proposed algorithm first transforms the original image into the DWT domain. Watermark is specially designed user information that is represented by images, text characters, sound data, and so on. In this paper, we used a watermark image because of the convenience of visual analysis and evaluation. The watermark is inserted in the lowest frequency region of the 3-level DWT. After inverse DWT of the watermarked DWT coefficients, an attack analysis key is inserted in the spatial domain. Figure 1 shows the overview of the proposed algorithm.

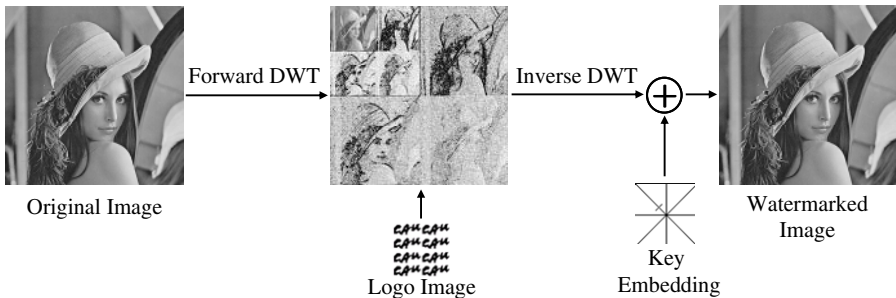


Fig. 1. Overview of watermark embedding procedure using wavelet filters

2.1 Wavelet Transform

Wavelet transform has been independently developed in many areas such as mathematics, electrical engineering, medical imaging, and communication. Especially, wavelet transform in JPEG2000 image compression standard provides high compression ratio and high image quality comparing to existing JPEG compression.

A single level DWT divides an image into 4 coefficient images. Each coefficient image contains one of low frequency bands and high frequency bands. With an $M \times N$ image, 2-D DWT generates four $M/2 \times N/2$ coefficients: LL, LH, HL, and HH, where LL represents a low frequency band, LH a horizontal high frequency band, HL vertical high frequency band, HH a diagonal high frequency band. The low frequency band is utilized to the next level of DWT. The sub-band structure of a 3-level DWT is shown in Figure 2. The watermark insertion in the proposed algorithm is performed in the LL_3 that is the lowest frequency band in the 3-level DWT.

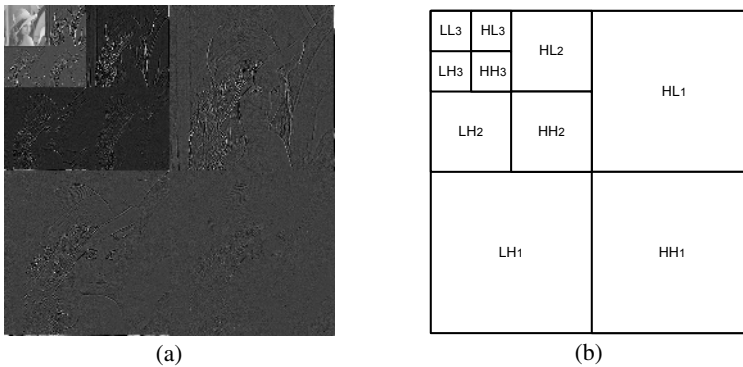


Fig. 2. (a) 3-Level wavelet decomposed image and (b) its convention

2.2 Watermark Embedding Algorithm

The proposed algorithm consists of a frequency-domain watermark insertion part using DWT and a spatial-domain key insertion part for attack analysis. The low frequency band remains robust to attacks. Thus, the watermark insertion in the proposed algorithm is carried out in the LL_3 band for a 3-level DWT. It should be noted that watermark insertion should be carefully designed since the coefficients of LL_3 band have strongest signal energy. Strong watermark could be visible at the LL_3 band. After inverse DWT of the watermarked DWT coefficients, a specially designed key is inserted in the middle of the image. The key helps to estimate the geometric transformation due to attacks. Figure 3 shows the block diagram of watermark and key insertion in the proposed algorithm.

The algorithm for embedding 2-bit gray scale logo is formulated as follows:

Step 1: Apply a 3-level DWT to an input original image $f(x,y)$ ($512 \times 512 \times 8$ bits) with the Daubechies D4 wavelet filters, which generates 9 subbands of high frequency ($LH_i, HL_i, HH_i, i=1 \sim 3$) and one low-frequency subband (LL_3).

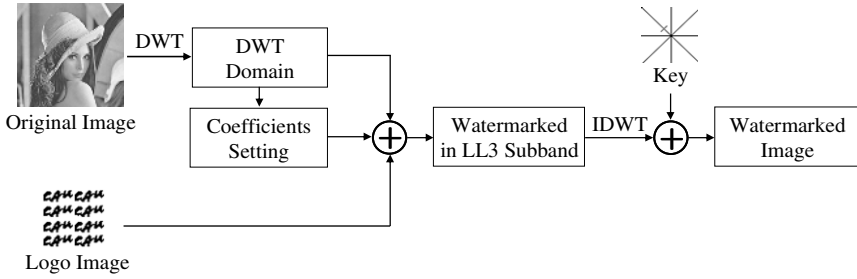


Fig. 3. The block diagram of watermark embedding procedure

Step 2: Check the magnitude of LL3 coefficients and find the location of bits where the 2-bit logo image will be inserted. Taking in to account the visual response curve (VRC) of human eye, the proposed method inserts bit patterns of watermark using a pre-specified threshold.

```

for( i, j is 0 to N/8 ) {
    if ( Cij >= T ) then {
        Cij >> 5, Cij << 5
        Cij = Cij + Wij × 24
    }
    Else {
        Cij >> 4, Cij << 4
        Cij = Cij + Wij × 23
    }
}

```

Where, C is a wavelet coefficient of original image,
T is coefficient threshold,
W is a logo image(W_{ij} ∈ {0,1}, i,j=0~63).

Step 3: After obtaining $f^{\sim}(x,y)$ using inverse DWT, insert a 64×64 3-bit key pattern shown in Figure 3 at center of $f^{\sim}(x,y)$.

2.3 Watermark Detection Algorithm

Watermark detection algorithm consists of the key extraction part in spatial domain and watermark extraction part in LL3 of DWT domain. Geometrical transformation is analyzed by checking the spatial domain key followed by image restoration processing. The watermark is, then, extracted in the LL3 of DWT coefficients. Dong et al. [12] proposed an image normalization technique to make the blind extraction robust. The proposed algorithm does not replace the existing techniques. On the contrary, the proposed algorithm in conjunction with existing techniques can further improve the robustness of watermarking. It should be noted that the performance improvement of the proposed algorithm with existing techniques will be evaluated in the future study.

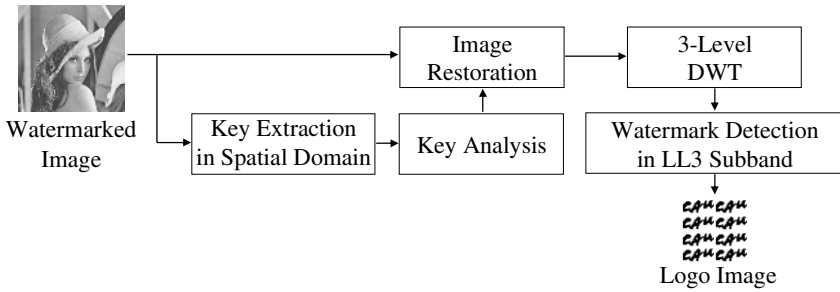


Fig. 4. The block diagram of watermark detection procedure

The detection algorithm shown in Figure 4 has the following procedures.

- Step 1: Find the key in the image modified by attacks and analyze geometrical attacks such as translation, scaling, etc. by comparing the extracted key with the original key.
- Step 2: Restore the modified image to the original watermarked image by reversing geometrical transformation found in step 1.
- Step 3: Apply the 3-level DWT to the restored image. The following pseudo codes represent the watermark extraction procedure in the LL3 band.

```

for( i, j is 0 to N/8 ) {
    if ( C'_{ij} >= T ) then {
        if C'_{ij} mod 2^5 > 2^4 + 2^2 then W'_{ij} = 1
        else W'_{ij} = 0
    }
    Else ( C'_{ij} < T ) {
        if C'_{ij} mod 2^4 > 2^3 + 2^1 then W'_{ij} = 1
        else W'_{ij} = 0
    }
}

```

*Where, C' is a wavelet coefficient of watermarked image,
T is coefficient threshold,
W' is a extracted watermark.*

3 Experimental Results

The performance of the proposed algorithm is tested on various types of images. The test image is a grayscale 8-bit Lena image whose size is 512×512. The logo used for watermarking is a 64×64 2-bit grayscale image. The original Lena image, a key image and a logo images are shown in Figure 5(a), (b) and (c) respectively. Daubechies D4 filter coefficients are used for 3-level wavelet decomposition. Performance evaluation of the proposed algorithm on test images with various characteristics was studied.

PSNR (Peak Signal-to-Noise Ratio) is used to analyze the quality of the watermarked image. PSNR is defined as

$$PSNR = 10 \log_{10} \frac{255^2}{\frac{1}{MN} \sum_{m,n} (I_{m,n} - I'_{m,n})^2} \tag{1}$$

where I represents the original image, I' the modified image and M and N represent image size. The number of mismatched data between the inserted watermark and the extracted watermark is used to represent the similarity of watermarks. NC (normalized correlation) for valid watermarks, which represents the characteristics of the extracted watermark, is defined as

$$NC = \frac{\sum_{x,y} w_{x,y} w'_{x,y}}{\sum_{x,y} w_{x,y}^2} \tag{2}$$

Where w represents the inserted watermark, w' the extracted watermark. The experimental results are rounded to the fourth decimal place. The NC for random noise is about 0.5 and possibility of distinguishing extracted logo more than 0.7~0.8 NC.

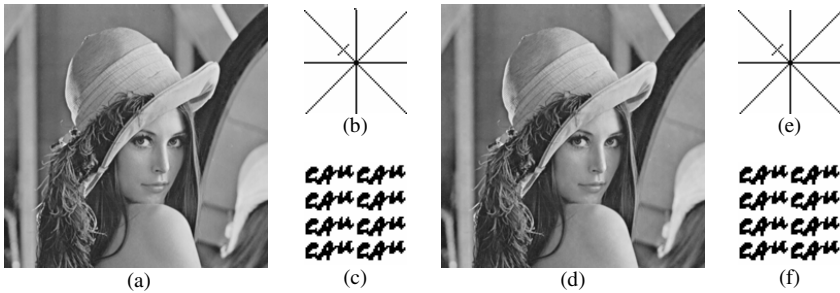


Fig. 5. (a)Original image, (b)key, (c)logo image, (d)watermarked image, (e)extracted key and (f)extracted logo image

The watermarked Lena image with PSNR 42.11db is shown in Figure 5(d). There is no perceptual degradation in the watermarked image. The extracted key and logo from the watermarked image are shown in Figure 5(e) and (f) respectively.

3.1 Non-geometric Attacks

We classify the attack patterns into non-geometric attacks and geometric attacks. Watermarked images are first tested for non-geometric attacks such as Gaussian filtering, median filtering and compression. The proposed algorithm can easily cope with wavelet-based JPEG2000 and DCT-based JPEG standard as shown in Table 1 and 2.

Table 1 shows the performance of the proposed algorithm against JPEG2000 compression attacks. JPEG2000 compression rates are set in between 0.1 and 0.45 bpp. The step size of the compression rate is 0.05. The extracted watermarks are seriously

Table 1. PSNRs and NCs of watermark embedded images at JPEG2000 compression

JPEG2000 Rate	Lena		Lake		Boat	
	PSNR	NC	PSNR	NC	PSNR	NC
0.45	-	0.998	48.13	0.974	48.13	0.999
0.40	48.13	0.997	45.12	0.972	48.13	0.973
0.35	48.13	0.968	43.36	0.961	45.12	0.978
0.30	45.12	0.969	41.14	0.973	42.11	0.894
0.25	43.36	0.968	39.68	0.858	40.35	0.892
0.20	41.14	0.903	37.34	0.773	38.59	0.857
0.15	39.68	0.854	35.58	0.692	36.99	0.731
0.10	37.72	0.773	33.36	0.601	34.91	0.667

JPEG2000 Rate	Goldhill		Drop		Peppers	
	PSNR	NC	PSNR	NC	PSNR	NC
0.45	48.13	0.999	-	1	-	0.998
0.40	48.13	0.977	-	1	48.13	0.990
0.35	45.12	0.981	-	0.998	45.12	0.971
0.30	41.14	0.885	-	0.997	45.12	0.975
0.25	41.14	0.885	48.13	0.994	42.11	0.908
0.20	37.34	0.796	46.93	0.937	40.35	0.871
0.15	37.34	0.796	45.12	0.923	38.59	0.814
0.10	34.91	0.681	43.36	0.894	36.99	0.718

Table 2. PSNRs and NCs of watermark embedded images at JPEG compression

JPEG Quality	Lena		Lake		Boat	
	PSNR	NC	PSNR	NC	PSNR	NC
10	43.36	1	42.11	1	42.11	1
8	39.68	0.998	36.67	0.999	37.72	0.992
6	39.10	0.988	35.58	0.989	36.99	0.991
4	36.67	0.915	33.51	0.917	34.71	0.917
2	34.15	0.832	30.97	0.830	31.80	0.823
0	32.11	0.622	29.27	0.614	30.00	0.622

JPEG Quality	Goldhill		Drop		Peppers	
	PSNR	NC	PSNR	NC	PSNR	NC
10	43.36	1	48.13	1	42.11	1
8	38.59	0.999	43.36	0.999	38.13	0.999
6	37.34	0.990	43.36	0.988	37.34	0.992
4	34.71	0.913	40.35	0.904	35.58	0.922
2	31.80	0.827	37.72	0.814	33.51	0.826
0	30.00	0.613	35.58	0.597	32.00	0.609

damaged below the rate 0.1. JPEG compression is also evaluated over various compression rates. Table 2 presents the PSNRs and NCs according to JPEG quality factors.

Table 3. PSNR and NC of watermark embedded images for different attacks

Attack	Lena		Lake		Boat	
	PSNR	NC	PSNR	NC	PSNR	NC
Gaussian	26.52	0.808	26.99	0.854	25.80	0.892
sharpening	18.67	0.735	15.13	0.639	16.23	0.671
Median(3×3)	26.23	0.873	26.28	0.811	25.19	0.854
Median(5×5)	22.40	0.738	21.81	0.710	21.27	0.696

Attack	Goldhill		Drop		Peppers	
	PSNR	NC	PSNR	NC	PSNR	NC
Gaussian	25.39	0.906	29.81	0.940	26.83	0.919
sharpening	17.31	0.658	20.56	0.815	16.59	0.727
Median(3×3)	24.95	0.873	29.50	0.927	26.49	0.882
Median(5×5)	21.57	0.692	24.29	0.828	21.60	0.760

We also include non-geometric attacks such as Gaussian, sharpening, and median filter. The PSNR of the attacked images are shown in Table 3. The attack patterns used in the experiment are from the Korean Watermarking Certification [13].

3.2 Geometric Attacks

The geometric attacks are further divided into translation, cropping, rotation, flip, and scale. The performances of key extraction and watermark extraction are individually

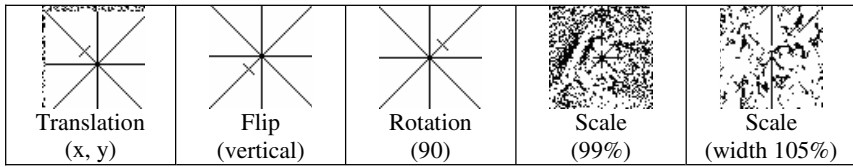


Fig. 6. Extracted key of attacked image

Table 4. Experimental results for the geometric attacks

geometric attacks	Lena	Lake	Boat	Goldhill	Drop	Peppers
Translation (0.5)	0.978	0.993	0.963	0.981	0.987	0.994
Translation (1)	0.988	0.987	0.989	0.992	0.989	0.988
Translation (10)	0.977	0.981	0.976	0.979	0.986	0.980
Translation (0.05%)	0.941	0.954	0.937	0.942	0.961	0.953
Flip(horizontal)	0.989	0.988	0.997	0.987	0.990	0.986
Flip(vertical)	0.981	0.976	0.986	0.992	0.995	0.972
Rotation(90)	0.988	0.965	0.968	0.979	0.986	0.963
Rotation(180)	0.970	0.975	0.984	0.973	0.984	0.986
Cropping(0.1)	0.903	0.891	0.908	0.846	0.829	0.834
Cropping(0.2)	0.839	0.827	0.837	0.768	0.752	0.760
Cropping(0.3)	0.791	0.656	0.802	0.669	0.654	0.659
Cropping(0.4)	0.746	0.567	0.751	0.578	0.573	0.572
Cropping(0.5)	0.727	0.512	0.723	0.524	0.518	0.517

tested. Once the key pattern is detected, the attacked image can be restored. The results of the key extraction are shown in Figure 6. The watermark extraction follows the key extraction.

The geometric parameters are evaluated to determine the geometric attack pattern, and the performance of the watermark extraction is listed in Table 4.

The proposed watermarking algorithm can endure translation attacks if the translation is more than 0.05%, horizontal or vertical flip attacks, and 90 degree rotation attacks. The watermark can be identified with cropping attacks up to 0.5%.

4 Conclusion

In this paper, we propose a novel blind watermarking technique that has a spatial domain key and a DWT domain watermark. The proposed technique utilizes a spatial domain key to analyze various attack patterns. By restoring the geometrical attacks, we can improve the robustness of the proposed watermarking algorithm. The DWT is also utilized to make the watermark survive various attacks.

We evaluate the performance of the proposed watermarking technique against various attacks including geometric transformation and image compression such as JPEG and JPEG2000. The experimental results show that the proposed technique has the improved performance compared to the conventional DWT-based watermarking technique.

Further research will include a study on an enhanced key pattern to make the proposed algorithm analyze and endure more complicated attacks.

Acknowledgment

This research was supported by Korea Ministry of Science and Technology under the National Research Laboratory. Project and by Korean Ministry of Information and Communication under HNRC-ITRC program at Chung-Ang University supervised by IITA.

References

1. Kang, X., Huang, J., Shi, Y., Lin, Y.: A DWT-DFT composite watermarking scheme robust to both affine transform and JPEG compression. *IEEE Trans. Circuits and Systems for Video Technology*, Vol. 13. (2003) 776-786
2. Pereira, S., Pun, T.: Robust template matching for affine resistant image watermarking. *IEEE Transactions on image processing*, Vol. 9. (2000) 1123-1129
3. Liu, H., Liu, J., Huang, J., Huang, D., Shi, Y. Q.: A robust DWT-based blind data hiding algorithm. *Circuits and Systems. ISCAS 2002. IEEE International Symposium on*, Vol. 2. (2002) 672-675
4. Reddy, A. A., Chatterji, B. N.: A new wavelet based logo-watermarking scheme. *Pattern Recognition Letters*, Vol. 26. (2005) 1019-1027
5. Dawei, Z. L., Guanrong, C., Wenbo, L.: A chaos-based robust wavelet-domain watermarking algorithm. *Chaos, Solitons & Fractals*, Vol. 22. (2004) 47-54

6. Kim, Y., Yoo, J., Lee, S., Shin, J., Paik, J., Jung, H.: Adaptive mode decision for H.264 encoder. *Electronics Letters*, Vol. 40. (2004) 1172-1173
7. Wiegand, T., Sullivan, G., Bjontegaard, G., Luthra, A.: Overview of the H.264/AVC video coding standard. *IEEE Trans. Circuit, System for Video Technology*, Vol. 13. (2003) 560-576
8. Deguillaume, F., Voloshynovskiy, S., Pun, T.: A method for the estimation and recovering from general affine transform in digital watermarking applications. *Proc. SPIE: Security and Watermarking of Multimedia Contents IV*, Vol. 4675. (2002) 313-322
9. Dugelay, J. L., Petitcolas, F. A. P.: Possible counter-attackers against random geometric distortions. *Proc. SPIE: Security and Watermarking of Multimedia Contents II*, Vol. 3971. (2002)
10. Braudaway, G. W., Minter, F.: Automatic recovery of invisible image watermarks from geometrically distorted images. *Proc. SPIE: Security and watermarking of multimedia contents I*, Vol. 3971. (2000)
11. Kang, X., Huang, J., Shi, Y.: An image watermarking algorithm robust to geometric distortion. *Lecture Notes in Computer Science: Proc. Int. Workshop on Digital Watermarking 2002*, Vol. 2613. (2002) 212-223
12. Dong, P., Brankov, J. G., Galatsanos, N. P., Yang, Y., Davoine, F.: Digital watermarking robust to geometric distortions. *IEEE Trans. on Image processing*, Vol. 14. (2005) 2140-2150
13. Oh, W., Kim, H.: The watermarking evaluation & certification technique of image. *Telecommunications Technology Association*, Vol. 90. (2003) 95-103

Probability Approximation Using Best-Tree Distribution for Skin Detection

Sanaa El Fkihi^{1,2}, Mohamed Daoudi¹, and Driss Aboutajdine²

¹ FOX-MIIRE LIFL (UMR CNRS-USTL 8022) Telecom Lille1, France
{elfkihi, daoudi}@enic.fr

² GSCM Faculty of Sciences Rabat, University Mohammed V, Morocco
aboutaj@fsr.ac.ma

Abstract. Skin detection consists in detecting human skin pixels from an image. In this paper we propose a new skin detection algorithm based on approximation of an image patch joint distribution, called Best-Tree distribution. A tree distribution model is more general than a bayesian network one. It can represent a joint distribution in an intuitive and efficient way. We assess the performance of our method on the Compaq database by measuring the Receiver Operating Characteristic curve and its under area. These measures have proved better performances of our model than the baseline one.

1 Introduction

Skin detection plays an important role in various applications such face detection [1], searching and filtering image content on the web [2], ... Research has been performed on the detection of human skin pixels in color images by the use of various statistical color models [3]. Some researchers have used skin color models such as Gaussian, Gaussian mixture or histograms [4]. In most experiments, skin pixels are acquired from a restricted number of people under a limited range of lighting conditions.

Unfortunately, the illumination conditions are often unknown in an arbitrary image, thus the variation in skin colors is lesser constrained in practice. However, given a large collection of labeled training pixels including all human skin (Caucasians, Africans, Asians, ...), we can still model the distribution of skin and non-skin colors in the color space.

Recently Jones and Rehg [5] proposed some techniques for skin color detection by estimating the distribution of skin and non-skin colors using labeled training data. The comparison results of histogram and Gaussian mixture density models estimated with EM algorithm found that the histogram models is slightly superior in terms of skin pixel classification performance for the standard 24-bit RGB color space.

Even if different criteria are used for evaluation, a skin detection system is never perfect. General appearance of the detected skin-zones, or other global criteria might be important for further processing.

For quantitative evaluation, we will use false positives and detection rates. False positive rate is the proportion of non-skin pixels classified as skin whereas detection rate is the proportion of skin pixels classified as skin. The user might wish to combine these two indicators in his own way depending on the percentage of error he can afford. Hence we propose a system in which the output is not binary but floating number between zero and one, where the larger value is considered as the larger belief of a skin pixel. Thus a user can apply a threshold to obtain a binary image, then the error rates for all possible thresholding will be summarized in the Receiver Operating Characteristic (ROC) curve.

The aim of this paper consists in learning the dependencies between the pixels within an image patch, to classify skin and non-skin textures. To achieve this goal, we draw the inspiration from a tree distribution method developed in [6], which we assume that is more general than a bayesian network [7].

The tree representation of an object allows an efficient search, and simple learning. In the case where the tree represents only one class, the learning problem solution is an algorithm developed by Chow and Liu[8]. This algorithm minimizes the Kullback-Leibler divergence [9] between the true distribution and the tree approximated distribution.

Our main contribution will be to construct one tree distribution representing two probability mass functions corresponding to two different classes.

The paper is organized as follows: in section 2, we introduce the notations that will be used throughout the paper, and present the features used. Section 3 details our tree distribution classifier model. Section 4 is devoted to experiments and comparisons with an alternative method. Finally, in section 5 conclusions and perspectives are drawn .

2 Notations and Methodology

In this section, we introduce the notations that will be used in this paper. We note s a pixel, and S the set of image pixels. (i_s, j_s) is the coordinate of s . We consider the RGB color space, the color of s is x_s . The "skinness" of a pixel s , is y_s with $y_s = 1$ if s is a skin pixel and $y_s = 0$ if not. The color image, which is the vector of color pixels, is notated x and the binary image made up of the y_s 's is notated y . In order to take into account the neighboring influence between pixels, we define the following neighborhood system :

$$V_s^r = \{(i, j) / |i - i_s| < r, |j - j_s| < r\} \setminus \{(i_s, j_s)\} \quad (1)$$

where the parameter r takes an integer value.

The figure 1 shows different orders neighborhood system in which s denotes the considered site and the gray boxes its neighbors. In Fig.1(a) the value of r is 1 and the cardinal of V_s^1 is $|V_s^1| = 0$. Fig.1(b) gives the second order neighborhood system. There are eight neighbors of s ($|V_s^2| = 8$). The third order neighborhood system is shown in Fig.1(c) and $|V_s^3| = 24$.

Thus, we consider a vector of observations X which stands for an image patch ($k \times k$, $k = 2r - 1$). The elements of the patch are decomposed until a low-level, the

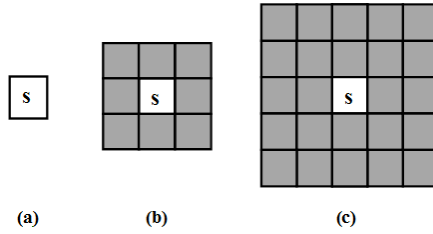


Fig. 1. The neighboring sites in the n -th order neighborhood system

resultant vector is $X = (x_1, x_2, \dots, x_n)$ where $n = 3k^2$. For each x_l component of X , V_l^r refers to the set of x_l neighbors ($l \in \{1, 2, \dots, n\}$).

Let us assume for a moment that we knew the joint probability distribution $Pro(X, y_s)$ of the vector (X, y_s) , then Bayesian analysis tells us that whatever cost function the user might think of, all what is needed is the a-posterior distribution $Pro(y_s|X)$.

From the user’s point of view, the useful information is contained in the one pixel marginal of the a-posterior probability, that is for each pixel, the quantity $Pro(y_s = 1|X)$ quantifies the skinness belief. In practice the model $Pro(X, y_s)$ is unknown, instead we have the segmented Compaq Database which it is a collection of samples $\{(x^{(1)}, y^{(1)}), \dots, (x^{(N)}, y^{(N)})\}$. For each $1 \leq i \leq N$, $x^{(i)}$ is a color image, and $y^{(i)}$ is its binary skinness image associated. We suppose that the samples are independent of each other with the distribution $Pro(X, y_s)$. The collection of samples will be referred as the training data.

Our objective is to construct a probabilistic classifiers that represent the posterior probabilities ($Pro(y_s = 1|X)$ and $Pro(y_s = 0|X)$) of skinness at pixel s given its neighbors, and using a single tree distribution.

In the next, we will use $Pro(X|y_s = 1) = p(X)$ and $Pro(X|y_s = 0) = q(X)$ to simplify notations.

To find a non-oriented acyclic graph (tree) modelling $Pro(X, y_s)$, we consider a non-oriented graph $G(V, E)$ corresponding to X . Each element x_u of X is viewed as a vertex ($x_u \in V$). The set of the edges E encloses all relationships between two elements of $V_l^r \cup \{(i_l, j_l)\}$ where $l = 1, 2, \dots, n$. Two neighbor vertices x_u and x_v are noted $u \sim v$.

3 The Tree Distribution Model

We consider a probabilistic classifier that represents the a-posteriori probability by using tree models. Thus, we suppose that the graph G is a tree: $G(V, E)$ is a connected graph without loops, noted T . In this case [10], we can proof that the probability approximated by the tree T is :

$$Pro_T(x) = \prod_{(u \sim v) \in T} \frac{Pro_{uv}(x_u, x_v)}{Pro_u(x_u)Pro_v(x_v)} \prod_{x_u \in V} Pro_u(x_u) \tag{2}$$

where $Pro_u(x_u)$ is one-vertex marginal of Pro and $Pro_{uv}(x_u, x_v)$ is its two-vertex marginal, defined as:

$$Pro_u(x_u) = \sum_{x_v \in V; v \neq u} Pro_T(x) \quad (3)$$

$$Pro_{uv}(x_u, x_v) = \sum_{x_u \in V; u \neq v} Pro_T(x) \quad (4)$$

$$\sum_{x_u \in V} Pro_T(x) = 1 \quad (5)$$

3.1 Learning of Tree Distribution

The learning problem is formulated as follows: given a set of observations $X = (x_1, x_2, \dots, x_n)$, we want to find one tree T in which the distribution probability is optimum for two different classes: skin and non skin. We mean by optimum that the distance between two probabilities of the classes is maximum, and we give the following definition:

Definition 1. *The Best-Tree skin classifier is a system based on a tree in which the distribution probability is optimum for Skin and NON-Skin classes.*

To deal with the learning problem, we propose to maximize the Kullback-Leibler divergence (KL) between two probability mass functions $p(X)$ and $q(X)$ corresponding to two different classes. Therefore, we give the following statement:

Statement 1. *Probability distributions of dependence tree, $p_T(\mathbf{x})$ and $q_T(\mathbf{x})$ are respectively the optimum approximations to the true probabilities $p(\mathbf{x})$ and $q(\mathbf{x})$ if and only if their dependence tree T has the maximum weight defined by :*

$$\sum_{(u \sim v) \in T} \{KL(p_{uv}, q_{uv}) - KL(p_u, q_u) - KL(p_v, q_v)\} \quad (6)$$

where p_u and q_u are respectively one-vertex marginal of p and q , and p_{uv} and q_{uv} are theirs two-vertex marginal.

The proof of our statement is postponed to the appendix 5.

In order to give detailed description of our model, we present the following procedure (1):

Procedure 1. *Best-Tree distribution processing*

- Input : Dataset $\{(x^{(1)}, y^{(1)}), \dots, (x^{(n)}, y^{(n)})\}$.
- Steps :

1. Fix $r \in \mathbb{N}$ to define the neighborhood system as :

$$V_s^r = \{(i, j) / |i - i_s| < r, |j - j_s| < r\} \setminus \{(i_s, j_s)\} \quad ; \quad s \in S$$

2. Consider a patch ($k \times k$, $k = 2r - 1$) centered in s . Determine V_b^r the set of the neighbors of each b in the patch. Decompose each element of V_b^r until a low-level and construct $V_{b^l}^r$ ($l = 1, 2, 3$), where b^l is a component of b in the RGB space.
3. Build X , where each element is obtained from $\bigcup_{b^l} V_{b^l}^r$.
4. Build a non-oriented graph $G(V, E)$ corresponding to X ; each element of X is a vertex. E is the set of edges corresponding to the relationship between two elements of V_l^r , ($l = 1, 2, \dots, n$).
5. Let x_u and x_v be two different vertices. Use the empirical estimators to compute the two-vertex marginal $p_{uv}(x_u, x_v)$ and $q_{uv}(x_u, x_v)$ of p_T and q_T as :

$$p_{uv}(x_u = i, x_v = j) = f_{ij}^1(x_u, x_v)$$

$$q_{uv}(x_u = i, x_v = j) = f_{ij}^0(x_u, x_v)$$

Where for $m \in \{1, 0\}$, $f_{ij}^m(x_u, x_v)$ is the sample joint frequency of $x_u = i$ and $x_v = j$ such as theirs labels are 1 or 0.

6. Compute the cost defined by the expression :

$$KL(p_{uv}, q_{uv}) - KL(p_u, q_u) - KL(p_v, q_v) \quad \forall u \sim v$$

7. Apply a Chow and Liu algorithm to build a maximum weighted spanning tree (MWST)[8].

– Output : Best-Tree distribution T .

3.2 Inference

We would like to compute the state of the pixel y_s given the observation vector X . By applying the Bayes' rule, we obtain:

$$Pro(y_s = j|X) = \frac{Pro(y_s = j)Pro(X|y_s = j)}{Pro(X)} \quad , \quad j = 0, 1. \quad (7)$$

Moreover,

$$Pro(X) = \sum_{y_s=0}^1 Pro(X, y_s) = \sum_{i=0}^1 Pro(X|y_s = i)Pro(y_s = i)$$

Where

$$Pro(X|y_s = 0) \approx q_T(X) = \prod_{(u \sim v) \in T} \frac{q_{uv}(x_u, x_v)}{q_u(x_u)q_v(x_v)} \prod_{x_u \in V} q_u(x_u) \quad (8)$$

$$Pro(X|y_s = 1) \approx p_T(X) = \prod_{(u \sim v) \in T} \frac{p_{uv}(x_u, x_v)}{p_u(x_u)p_v(x_v)} \prod_{x_u \in V} p_u(x_u) \quad (9)$$

$$Pro(y_s = 0) \approx q_T(y_s = 0) \quad (10)$$

$$Pro(y_s = 1) \approx p_T(y_s = 1) \quad (11)$$

All the elements of eq. (8), eq. (9), eq. (10), and eq. (11) are previously computed in step (5) of our processing (procedure 1).

4 Skin Detection Experiments

All experiments are made by using the following protocol. The Compaq database [5] contains about 18,696 photographs. It is randomly split into two almost equal parts. The first part, containing nearly 2 billion pixels, is used as training data; while the other one, the test set, is left aside for ROC curve and the Area Under the Curve (AUC) computations.

In our skin detection application we consider a (3×3) image patch. However, we use RGB color space, therefore the size of the observation vector X is 27. The Compaq Database is large enough; so the crude histograms, made with 512 color value per bin uniformly distributed, do not over-fit. Each histogram is then made of 32 bins. Experiments with this model are presented in figures 2 and 3.



Fig. 2. Best-Tree distribution model inputs and outputs

The top of figure 2 shows the original color images, and its bottom represents the result of our skin detection model.

The curves of figure 3 compare the performance of the Baseline model, which is an independent model[4], with the Best-Tree distribution performance. The x-axis represents the false positive rate, while y-axis corresponds to the true positive rate (the detection rate). The Baseline model is shown with blue triangles and the Best-Tree distribution with red stars.

Bulk results in the ROC curve of Figure 3 show an improvement of the performance around 6.5%. At 6.5% of false positive rate, the Baseline permits to detect 65.5% of skin pixels while the Best-Tree distribution model detects 71.8%.

Figure 4 shows some cases where our detector failed. The first row represents the original images, and the second one shows their corresponding skin maps obtained by the Best-Tree distribution. In the two first columns, examples of the non-skin pixels detected as skin pixels are given; succeeded by the skin pixels

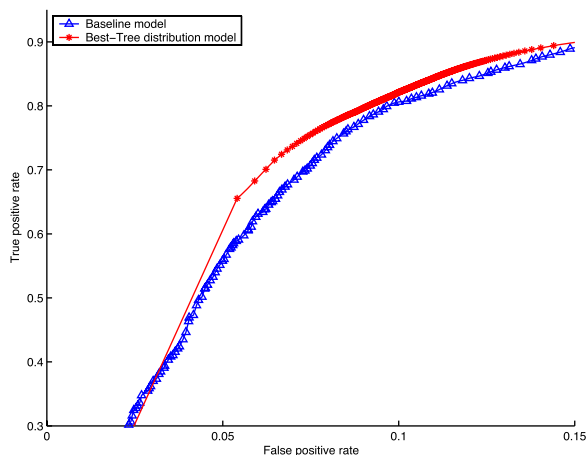


Fig. 3. The Baseline and the Best-Tree distribution ROC curves

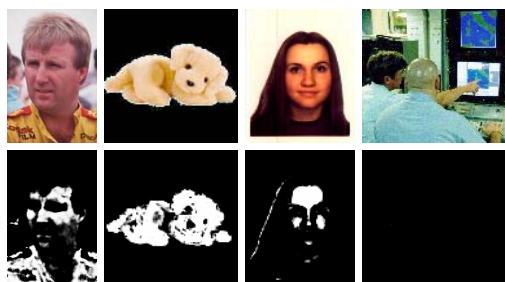


Fig. 4. Examples where the Best-Tree distribution model fails

detected as a non-skin pixels examples. This fail is due to over-exposure, or to skin-like color.

Another way to compare classification algorithms over multiple thresholding values is to compute the area under the ROC curve (AUC). Using $[0; 0.15]$ for integration interval, the normalized AUC, which equals to 0.0539 for the Baseline model and 0.1203 for our approach, confirms the obtained results for a single false positive rate.

5 Conclusion

In this paper, we have presented a new skin detection algorithm based on approximation of the image patch joint distribution. By making some assumptions, we propose The Best-Tree distribution model which maximizes a Kullback-Leibler divergence between two different probability distributions of classes : skin and non skin. Performance measured by the ROC curve on the Compaq database

shows an increase in detection rate from 3% to 15% for the same false positive rate of the Best-Tree distribution compared to the Baseline model, furthermore the AUC measures prove that.

In further work, we propose to apply the Best-Tree distribution for skin detection to block the Web adult images. Moreover, we constat that our approach could be used to classify other binary textures.

Acknowledgement

This work is supported by Maroc Telecom under the project "Filtrage de Sites à Contenu Illicite sur Internet", N° 105 908 05/PI. The authors would like to thank Prof. Bruno Jedynek of Johns Hopkins University for his help to this research.

References

1. Terrillon, J.C., Shirazi, M.N., Fukamachi, H., Akamatsu, S.: Comparative performance of different skin chrominance models and chrominance spaces for the automatic detection of human faces in color images. In: Fourth International Conference On Automatic Face and gesture Recognition. (2000) 54–61
2. Zheng, H., M.Daoudi, Jedynek, B.: Blocking adult images based on statistical skin detection. *Electronic Letters on Computer Vision and Image Analysis* **4** (2004) 1–14
3. Jedynek, B., Zheng, H., M.Daoudi: Skin detection using pairwise models. *Image and Vision Computing* **23** (2005) 1122–1130
4. Jones, M.J., Rehg, J.M.: Statistical color models with application to skin detection. Technical Report CRL 98/11, Compaq (1998)
5. Jones, M., Rehg, J.M.: Statistical color models with application to skin detection. In: *Computer Vision and Pattern Recognition*. (1999) 274–280
6. Meila, M., Jordan, M.I.: Learning with mixtures of trees. *Journal of Machine Learning Research* **1** (2000) 1–48
7. Sebe, N., Cohen, I., Huang, T.S., Gevers, T.: Skin detection : A bayesian network approach. (2004)
8. Chow, C.K., Liu, C.N.: Approximating discrete probability distributions with dependence trees. *IEEE Transactions on Information Theory* (1968) 462–467
9. Cover, T.M., Thomas, J.A.: *Elements of Information Theory*. Wiley-Interscience (1991)
10. Pearl, J.: *Probabilistic Reasoning in intelligent systems: networks of plausible inference*. Morgan Kaufmann (1988)

Appendix: Proof of the Statement 1

Proof. We assume that it exists a tree T in which each vertex is an X variable; T models two distributions p and q respectively approximated by p_T and q_T referenced on eq.(2).

The Kullback-Leibler divergence between p_T and q_T is :

$$\begin{aligned}
 KL(p_T, q_T) &= \sum_{x \in V} p_T(x) \log \frac{p_T(x)}{q_T(x)} \\
 &= \sum_{x \in V} p_T(x) \log p_T(x) - \sum_{x \in V} p_T(x) \log q_T(x) \tag{12}
 \end{aligned}$$

$$\begin{aligned}
 \sum_{x \in V} p_T(x) \log p_T(x) &= \sum_{x \in V} p_T(x) \sum_{x_u \in V} \log p_u(x_u) \\
 + \sum_{x \in V} p_T(x) \sum_{(u \sim v) \in T} &(\log p_{uv}(x_u, x_v) - \log p_u(x_u)p_v(x_v)) \tag{13}
 \end{aligned}$$

$$\begin{aligned}
 \sum_{x \in V} p_T(x) \log q_T(x) &= \sum_{x \in V} p_T(x) \sum_{x_u \in V} \log q_u(x_u) \\
 + \sum_{x \in V} p_T(x) \sum_{(u \sim v) \in T} &(\log q_{uv}(x_u, x_v) - \log q_u(x_u)q_v(x_v)) \tag{14}
 \end{aligned}$$

From (13) - (14), Eq. (12) becomes

$$\begin{aligned}
 KL(p_T, q_T) &= \sum_{x \in V} p_T(x) \sum_{x_u \in V} \log \frac{p_u(x_u)}{q_u(x_u)} \\
 + \sum_{x \in V} p_T(x) \sum_{(u \sim v) \in T} &(\log \frac{p_{uv}(x_u, x_v)}{q_{uv}(x_u, x_v)} - \log \frac{p_u(x_u)p_v(x_v)}{q_u(x_u)q_v(x_v)}) \tag{15}
 \end{aligned}$$

Thus, we obtain :

$$\begin{aligned}
 KL(p_T, q_T) &= \sum_{x_u \in V} KL(p_u, q_u) + \\
 \sum_{(u \sim v) \in T} &\{KL(p_{uv}, q_{uv}) - KL(p_u, q_u) - KL(p_v, q_v)\} \tag{16}
 \end{aligned}$$

Since, for all $x_u \in V$, the $KL(p_u, q_u)$ are independent of the dependence tree, and KL divergence is non-negative. Maximizing the closeness measure $KL(p_T, q_T)$ is equivalent to maximizing the total branch weight eq. (6).

Fusion Method of Fingerprint Quality Evaluation: From the Local Gabor Feature to the Global Spatial-Frequency Structures

Decong Yu¹, Lihong Ma^{1,2}, Hanqing Lu², and Zhiqing Chen³

¹ GD Key Lab. of Computer Network, Dept. of Electronic Engineering, South China Univ. of Tech., Guangzhou, China, 510640

yudecong@163.com, eelhma@scut.edu.cn

² National Lab of Pattern Recognition, Inst. Automation, Chinese Academy of Science, Beijing, China 100080

luhq@nlpr.ia.ac.cn

³ Criminal Tech. Center, Dept. of Public Security of Guangdong Province, China 510050

zq_chen@163.com

Abstract. We propose a new fusion method to evaluate fingerprint quality by combining both spatial and frequency features of a fingerprint image. In frequency domain, a ring structure of DFT magnitude and directional Gabor features are applied. In spatial domain, black pixel ratio of central area is taken into account. These three features are the most efficient indexes for fingerprint quality assessment. Though additional features could be introduced, their slight improvement in performance will be traded off with complexity and computational load to some extent. Thus in this paper, each of the three features are first employed to assess fingerprint quality, their evaluation performance are also discussed. Then the suggested fusion approach of the three features is presented to obtain the final quality scores. We test the fusion method in our public security fingerprint database. Experimental results demonstrate that the proposed scheme can estimate the quality of fingerprint images accurately. It provides a feasible rejection of poor fingerprint images before they are presented to the fingerprint recognition system for feature extraction and matching.

1 Introduction

Fingerprint recognition system is widely used in criminal identification, ATM card verification and access control, due to its feature's individual uniqueness and age invariability. But the performance of an Automatic Fingerprint Identification System (AFIS) depends heavily on fingerprint quality which mainly concerned with skin humidity, impressing pressure, dirt, sensing mechanism, scar and other factors. A fingerprint of good quality should have clear ridge and valley patterns, and could guarantee a high performance of recognition. Therefore, an efficient criterion for fingerprint quality evaluation will be of benefit to practical applications, such as quality control of fingerprint acquisition, quality distribution analysis of fingerprint database, and threshold decision for modification of low quality images. If the fingerprint

quality is assessed at first, the images of poor quality could be discarded and the fingerprint acquisition repeated, the AFIS performance will finally be greatly improved.

Some methods have been proposed to evaluate fingerprint quality in the past few years. Hong et al [1] quantified the quality of a fingerprint image by measuring the block variance of gray levels, which was computed in directions orthogonal to the orientation field. The variance was then used to decide the fingerprint quality in terms of the contrast of a considered block. However, this method had to be carried out in a precise orientation field which may not be correctly obtained in heavy noise situation. In addition, it is computational expensive. Ratha and Bolle [2] proposed another method for quality estimation in wavelet domain for Wavelet Scalar Quantization (WSQ) images. But WSQ is not a necessary step for uncompressed fingerprint images in AFIS. Shen et al [3] described a Gabor feature based approach to quality measurement, it also suffers from the parameter setting and excessive computation load of Gabor transform. Lim et al [4] employed the ratio of eigenvalues of the gradient vectors to estimate local ridge and orientation certainty, and determined the quality with the orientation flow. This can indicate the confidence of orientation estimation, provided the noise is not directional distributed. Other quality assessments include the Fourier spectrum based method [5] and the gradient oriented scheme [6], all these methods utilized only the partial information which is not sufficient to measure a fingerprint image. Hence, Global and local information should be combined together to accomplish the evaluation task. A hybrid method joined seven local and global features of fingerprint [7] to assess quality was reported in 2005. But it is difficult to balance the quality weight of each feature, its linear weighting does not regard the nonlinear contributions of some features.

In this paper, our research is focused on an accurate and feasible method for fingerprint quality measurement. A new nonlinear fusion method for quality scoring is suggested based on the combination of three efficient quality features: ring structures in frequency spectrum, Gabor features denoted directional information, and the black pixel ratio of central region which reflects the contrast and the integrity of the ridge and valley patterns. Since the central region around a core point is vital to fingerprint quality evaluation, all the calculation is performed on the central region of a core with the background region removing.

The remainder of this paper is organized as follows. Different quality features and their evaluation performance are presented in Section 2. The new fusion criterion and the quality scoring principle are proposed in Section 3. Experimental results are shown in Section 4 to demonstrate the validness of our method. Finally, the conclusions and discussions are given in Section 5.

2 Features and Evaluation

The fingerprint quality could be examined by many features. Three most important ones to quantify the fingerprint image quality include black pixel ratio of central area, ring structure of frequency spectrum, and directional Gabor Feature. In this section, we define three quality scoring functions with corresponding respect to these three features. The quality evaluation performances of each feature are compared as well.

2.1 Black Pixel Ratio of Central Area

Gray level distribution of the region around a core point is an essential index for fingerprint quality evaluation. Fingerprint images with high contrast relate to the well separated ridges and valleys. The smearing of wet fingerprint, the disconnected ridges in dry images and the background pixels introduced in a fingerprint will bias the ridge-valley contrast, thus a good quality image will have high contrast between ridges and valleys, while a poor quality image has low pattern contrast, and its ridge-valley structures are usually corrupted to some extent.

To quantify the ridge-valley contrast, let the black pixel ratio be R_B and the number of black pixels be N_B , w is the region length and width.

$$R_B = \frac{N_B}{w^2} \tag{1}$$

We suggest that a quality score Q_I can be calculated as follows:

$$Q_I = \begin{cases} \left(2 - \frac{G_{mean}}{150}\right) \left(1 - \frac{1 - R_B}{3}\right) & G_{mean} > Th_h \\ 1 - \sqrt{|R_B - 0.6|} & Th_l \leq G_{mean} \leq Th_h \\ \left(\frac{G_{mean}}{150}\right) \left(1 - \frac{R_B}{3}\right) & G_{mean} < Th_l \end{cases} \tag{2}$$

Where G_{ij} is the intensity of pixel (i,j) . G_{mean} denotes the average intensity of the 320×320 central region with $w = 320$ is the region length and width.

$$G_{mean} = \frac{1}{w^2} \sum_{i=1}^w \sum_{j=1}^w G_{ij} \tag{3}$$

For comparison, the original image and its central region are shown in Fig.1.

This scoring function is derived from R_B statistical analysis of the three categories fingerprint images as shown in Fig.2. We could observe from Fig.2 that a wet image with a lower R_B , the better quality is. In contrast, the lower the black pixel ratio is, and the worse a dry image will be. On the other hand, good quality fingerprint has an around constant R_B and G_{mean} is valued between Th_l and Th_h . Th_l and Th_h are determined empirically. When R_B is far from the constant, the fingerprint quality declines quickly. Based on the above ideas, the quality score Q_I is given as the afore mentioned equation (2).

The steps of fingerprint image quality evaluation by Q_I are described below.

- 1) Locate the core point of a fingerprint image [8]. If no core point was detected, the image center is taken as a core point.
- 2) Compute the average gray level of the 320×320 central region, whose center is located at the core point.

- 3) Classify the fingerprint images into three categories based on G_{mean} automatically: wet, good, and dry. To each category, different thresholds are used to binarize the 320×320 region.
- 4) Compute the quality score Q_I using equation (2).

Q_I can detect dry and wet fingerprint images easily and the scores of good quality fingerprints are ranged in a certain extent. However, it fails to measure the quality of image whose fingerprint area is far less than 320×320 pixels, because in such case the 320×320 central area consists of background pixels whose contrast are much smaller than the foreground pixels.



Fig. 1. Central region segmentation (a) the original fingerprint image (640×640); (b) the central region (320×320)

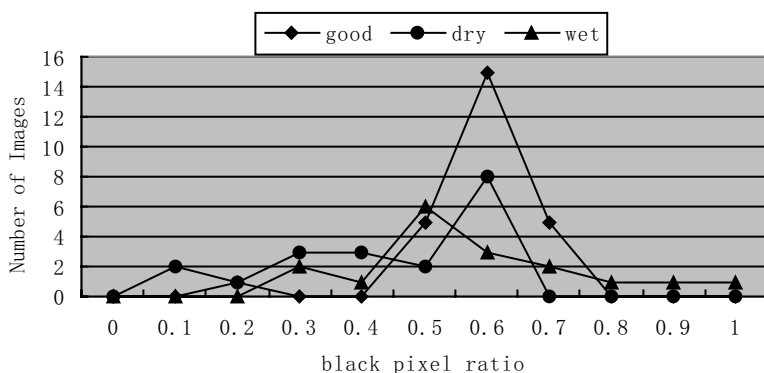


Fig. 2. The black pixel ratio of the three categories (good, dry and wet) fingerprint image

2.2 Ring Structure in Frequency Domain

Fourier transform is a useful analysis tool for distinguished global directional and frequent structures. Since ridges and valleys of fingerprint appears a distinct pattern in frequency domain, it could be applied to illustrate the fingerprint quality.

The DFT image of a good quality fingerprint image shows a ring around the origin of the frequency coordinate, because the ridge-valley patterns are quasi-periodic structures and presents a dominant frequency in most directions with an almost uniform modulus. In contrast, fingerprint images of bad quality do not appear an obvious ring in the spectrum plane, for they contain smear points, blurred edges, disconnected ridges and so on which occupy a wide range of frequency.

Making use of the above frequency characteristic of ridge-valley structures, we could modify the scoring function Q_2 firstly defined [5]:

$$Q_2 = \frac{1}{C_2} \sum_{\theta=0}^{359} F_{\theta}. \quad (4)$$

where C_2 is a constant normalizing the quality score in the range [0,1].

$$F_{\theta} = 2P_{\theta}(x_m, y_m) - P_{\theta}(x_s, y_s) - P_{\theta}(x_l, y_l). \quad (5)$$

$P_{\theta}(x_m, y_m)$ denotes the largest ring along the angle θ in magnitude spectrum, and $(x_s, y_s) = \frac{1}{2}(x_m, y_m)$, and $(x_l, y_l) = \frac{3}{2}(x_m, y_m)$ are the reference points along the same direction.

In this paper, we detect the maximum spectrum ring first to retain an accurate quality score before the calculation of equation (4). This modification is based on the fact that the variation of ridge-valley distance in different fingerprints will result in a negative F_{θ} if we compute equation (5) directly. The detection is performed by selecting the i^{th} ring with the following equation:

$$i = \left\{ i, \quad \text{Max}_i \left\{ \sum_{\theta=0}^{359} P_{\theta}(x, y) \right\}, \quad 30 \leq i \leq 40. \right. \quad (6)$$

where $P_{\theta}(x, y)$ denotes the frequency band, and (x, y) is its coordinate. The integer i is the radius of the maximum spectrum ring.

The $P_{\theta}(x, y)$ band located at 30 to 40 pixels from the origin along the angle θ , because the central area of the fingerprint image is 320×320 region of interest, its average ridge and valley distance is around 9 pixels, thus the corresponding ring of high spectrum magnitudes will appear in the frequency band located at 30 to 40 pixels away from the origin in the spectrum image.

Our algorithm can be briefly stated as follows:

- 1) Locate the core point of a fingerprint image.
- 2) Perform FFT on the 320×320 central region.
- 3) Find the largest ring in frequency plane using equation (6).
- 4) Compute the score Q_2 with regard to the largest spectrum ring using eq. (4).

Since Fourier transform is calculated on the whole central region, Q_2 is a global estimate to fingerprint quality. Even if the fingerprint is partially bad, its quality score

remains low. This global frequency feature takes on advantage of accurate assessment of the whole region, but it lacks of the characteristic of local pattern measurement.

2.3 Directional Gabor Features

Gabor filter has both orientation and frequency selective properties, Fingerprint images of good quality have a strong orientation tendency and a well-defined spatial frequency. For blocks of good quality, Gabor feature of one direction or several angles are larger than those in others direction; while for bad quality blocks, the Gabor features become close to each directions. Hence the standard deviation of Gabor features of different orientations can be used to judge the fingerprint quality.

The general form of a 2D Gabor filter is defined by

$$h(x, y, \theta_k, f_0) = \exp\left\{-\frac{1}{2}\left(\frac{x_{\theta_k}^2}{\sigma_x^2} + \frac{y_{\theta_k}^2}{\sigma_y^2}\right)\right\} \cos(2\pi f_0 x_{\theta_k}) \quad k=1, \dots, m. \tag{7}$$

$$x_{\theta_k} = x \cos \theta_k + y \sin \theta_k. \tag{8}$$

$$y_{\theta_k} = -x \sin \theta_k + y \cos \theta_k. \tag{9}$$

where θ_k is the k^{th} orientation of the filter bank $h(x, y, \theta_k, f_0)$, f_0 is the frequency of a sinusoidal plane wave, m denotes the number of orientations, and σ_x and σ_y are the standard deviation of the Gaussian envelope along the x and y axes, respectively.

The magnitude Gabor feature at each $w \times w$ block centered at (X, Y) can be defined as:

$$g(X, Y, \theta_k, f_0) = \left| \sum_{x=-w/2}^{w/2-1} \sum_{y=-w/2}^{w/2-1} I(X+x, Y+y) h(x, y, \theta_k, f_0) \right|. \tag{10}$$

where $I(x, y)$ denotes the intensity of the pixel (x, y) , w is the size of a block. $\theta_k = \pi(k-1)/m, k = 1, \dots, m$. In our study, we still mainly focus on the central 320×320 region.

The blocked standard deviation of Gabor feature G is calculated as follows:

$$G = \left(\frac{1}{m-1} \sum_{k=1}^m (g_{\theta_k} - \overline{g_{\theta}})^2 \right)^{1/2}, \quad \overline{g_{\theta}} = \frac{1}{m} \sum_{k=1}^m g_{\theta_k}. \tag{11}$$

Since the central region consists of only the foreground blocks, we needn't perform fingerprint segmentation before Gabor filtering. The quality score of a fingerprint can be computed by summing the standard deviation of all the blocks.

$$Q_3 = \frac{1}{C_3} \sum_{i=1}^N G(i) . \tag{12}$$

where C_3 is also a normalizing constant ranging the quality score form 0 to 1.

In summary, Q_3 is aimed at assessing the orientation properties of fingerprint, strong orientation relates to good quality. The quality estimation by Q_3 can also be depicted as follows:

- 1) Locate the core point of the fingerprint image.
- 2) Divide the 320×320 central region into N blocks of size $w \times w$.
- 3) For each block centered at pixel (i,j) , compute the m Gabor features and standard deviation value G by equations (10) and (11) respectively.
- 4) Obtain the quality score Q_3 by summing standard deviation value G of all the blocks using equation (12).

3 Fusion Criterion and Quality Scoring

In this section, we will present a novel fusion criterion combined the three evaluation results mentioned before. Since each method has its advantages and drawbacks, our research mainly aims at finding an optimal fusion criterion making benefits of the above assessment.

We define a fusion criterion which calculates a quality score Q according to:

$$Q = \frac{1}{C} \sum_{i=1}^3 w_i Q_i^{k_i} , k_i = 1,2 . \tag{13}$$

where C is a normalizing constant which ranges the quality score in $[0,1]$, w_i denotes the weight of each quality score Q_i , and k_i is a power factor of each quality score Q_i .

The above three features contribute differently to fingerprint quality evaluation. The fingerprint images of three categories are labeled as two kinds good quality and bad quality images merging the original wet and dry fingerprints. The values of Fourier and Gabor features are large to good fingerprint images. While, the value of R_B of central region is abnormal to bad fingerprint images, Q_1 can efficiently detect the bad quality images, so the weight of Q_1 will be bigger to bad quality images than others, and k_1, k_2 and k_3 are set to 1,2 and 2 respectively, while Q_2 and Q_3 can quantify the good quality images very well, the weights of these two scores are bigger than Q_1 , and k_1, k_2, k_3 are set to 2,1 and 1 respectively. Based on the above analysis, the fusion quality score Q is defined as equation (13).

As the contribution of each feature to the final quality score is nonlinear, the above equation can perform a better classification of good and bad images than the linear method, which will be given in section 4.

4 Experimental Results

The fingerprint database used in this experiment consists of 62 fingerprint images. The size of each fingerprint image is 640×640 pixels with the resolution of 500dpi

and 256 gray levels. And our research focuses on the 320×320 central region, the size of each block is 16×16 pixels. We verify the evaluation performance of our method on this public security fingerprint database using black pixel ratio of central area, directional Gabor feature and ring structures of spectrum. Fig. 3 shows the image score distribution of the three different feature evaluation methods and the score distribution of the fusion approach.

As shown in Fig. 3, the Gabor feature method and Fourier spectrum method can easily classify the fingerprint images into two groups: good and bad quality. The central area black white pixel ratios method is able to detect bad fingerprint images whose black pixel ratio of central area is abnormal. To some images, though the Gabor feature score and Fourier spectrum score are high, the central area score is low. Under this similar circumstance, the fusion method is needed for getting a reasonable score.

Fig. 3 demonstrates that the fusion method can find the poor quality images easily. After observing the images, we find that those fingerprint images whose final score are less than 0.1 have very bad quality. In order to test our proposed fusion method, we have sent these images to fingerprint classification system [8]. With ten percents fingerprint images of poor quality rejected, the five-classification (whorl, right loop, left loop, arch and tented arch) accuracy can be increased from 90.6 percent to 94.3 percent.

Fig. 4 show the score distribution of our nonlinear fusion method and the linear method [7]. From the figure, we can find that nonlinear method show better performance of distinguishing good and bad fingerprint images as described in section 3. In order to compare with the results of linear method, we have also sent these images to fingerprint classification system [8]. Based on the linear method, we reject ten percents fingerprint images of poor quality, the five-classification accuracy can be increased from 90.6 percent to 92.5 percent. We can get the conclusion that fingerprint classification based on our fusion method shows 1.8 percent better classification accuracy than the linear method.

By rejecting 25 percent images of bad quality according to our fusion method of fingerprint quality evaluation, the fingerprint classification accuracy will be increased to 96 percent.

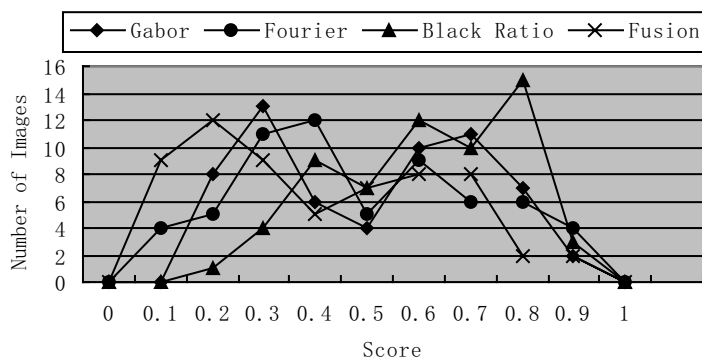


Fig. 3. Fingerprint image quality score distribution of different methods

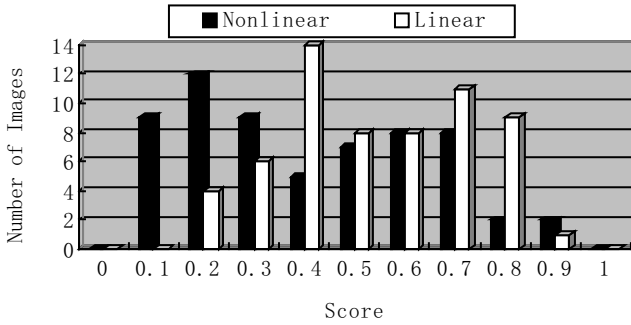


Fig. 4. Fingerprint image quality score distribution of nonlinear and linear methods

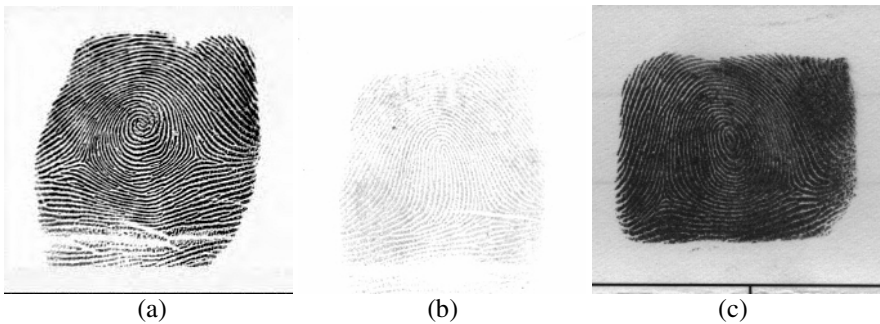


Fig. 5. Typical Fingerprint images: (a) good fingerprint; (b) dry fingerprint; (c) wet fingerprint

Fig.5 shows three typical fingerprint images. Based on our fusion method of fingerprint quality evaluation, fingerprint Fig.5 (a) whose contrast between ridges and valleys is high has the best score of all the fingerprints, fingerprint Fig.5 (b) with largely corrupted ridge structure and low contrast between ridges and valleys is the worst of all the dry fingerprints, and fingerprint Fig.5 (c) with low gray-level and low contrast ridges and valleys has the lowest score of the wet fingerprints.

5 Conclusions

We have developed a fusion method combining three features for quality evaluation of fingerprint image, which gives better performance than linear methods. Our experimental results demonstrate that the three features were sufficient for detecting poor quality fingerprint images. However, the proposed method relies on the correctly located core point and the foreground central 320×320 region. Further researches will emphasize on a more accurate core point detection for fusion method to improve the performance of the fingerprint image quality evaluation.

Acknowledgement

We would like to acknowledge the supports of China NNSF of excellent Youth (60325310), NNSF (60472063)& GDNSF/GDCNLF (04020074/CN200402).

References

1. Lin Hong, Yifei Wan, and Anil Jain, "Fingerprint image enhancement: algorithm and performance evaluation," IEEE Trans. Pattern Analysis Machine Intelligent, Vol.20, No. 8, pp. 777-789, August 1998
2. Nalini K. Ratha and R. Bolle, "Fingerprint image quality estimation," ACCV, PP.819-823, 2000
3. L.L. Shen, A. Kot and W.M. Koo, "Quality measure of fingerprint images," Third International Conference, AVBPA 2001, Halmstad, Sweden, Proceedings, pp. 266-271, Jun, 2001
4. E. Lim, X. Jiang, W. Yau, "Fingerprint quality and validity analysis", IEEE ICIP, 2002
5. Bongku Lee, Jihyun Moon, and Hakil Kim, "A novel measure of fingerprint image quality using Fourier spectrum" Proc. SPIE Vol. 5779,105(2005)
6. Jin Qi, Zhongchao Shi, Xuying Zhao and Yangsheng Wang, "Measuring fingerprint image quality using gradient" Proc. SPIE Vol. 5779,455(2005)
7. Qi, J., Abdurrachim, D., Li, D., Kunieda, H , "A hybrid method for fingerprint image quality calculation", Automatic Identification Advanced Technologies, 2005. Fourth IEEE Workshop on 17-18 Oct. 2005 Page(s):124 – 129
8. A.K. Jain, S. Prabhakar and H. Lin, "A Multichannel Approach to Fingerprint Classification," IEEE Trans. Pattern Analysis and Machine Intelligence, vol. 21, no. 4, pp. 348-359, 1999.

3D Face Recognition Based on Non-iterative Registration and Single B-Spline Patch Modelling Techniques

Yi Song and Li Bai

School of Computer Science and Information Technology
University of Nottingham, Jubilee Campus, Wollaton Road,
Nottingham NG8 1BB, UK
{yzs, bai}@cs.nott.ac.uk

Abstract. This paper presents a new approach to automatic 3D face recognition using a model-based approach. This work uses real 3D dense point cloud data acquired with a scanner using a stereo photogrammetry technique. Since the point clouds are in varied orientations, by applying a non-iterative registration method, we automatically transform each point cloud to a canonical position. Unlike the iterative ICP algorithm, our non-iterative registration process is scale invariant. An efficient B-spline surface-fitting technique is developed to represent 3D faces in a way that allows efficient surface comparison. This is based on a novel knot vector standardisation algorithm which allow a single B-Spline surface to be fitted onto a complex object represented as a unstructured points cloud. Consequently, dense correspondences across objects are established. Several experiments have been conducted and 91% recognition rate can be achieved.

1 Introduction

Recent theoretical and technical advance in 3D data capture opens up the possibility of 3D face recognition to overcome the difficulties in 2D face recognition systems, e.g. pose and illumination variations, as the 3D shape of a facial surface represents the anatomical structure of a face rather than the appearance. Whereas most of previous works use 2.5D face images (range data) [1,2,3,4], this work uses real 3D data acquired through a scanner based on the stereo photogrammetry technique, which captures the full frontal face in a single scan. However, 3D data (dense point clouds in this case) cannot be used directly for object recognition or shape analysis. First, the objects are in varied orientations and sizes. Second, the surface captured varies significantly across subjects and often includes neck or shoulders. There are often holes in the point clouds. Third, a 3D scan has about 30,000 vertices. So it is not very feasible to match a probe scan to every scan in the database using the Iterative Closest Point algorithm (ICP) [5,6].

Although ICP is a widely accepted method of registering unstructured point clouds without prior knowledge about topology, its scale and shape sensitivity make it impractical for face recognition. Thus, one of the motivations of this research is to explore a new registration method for 3D face recognition, which is scale and shape

invariant. On the other hand, how to establish dense correspondences across objects in an efficient and automatic way is another main motivation driving our research into investigating efficient 3D representation methods.

Besides an efficient registration method aiming for face recognition, the contribution of this paper also includes a new approach to face recognition based on 3D modelling which provides: 1) automatic dense correspondences establishment, 2) compact data representation.

The paper is organised as follows. In Section 2, related works are briefly reviewed. Section 3 describes our algorithm of scale invariant pose estimation. Section 4 presents an efficient single B-spline surface reconstruction method, based on which dense correspondences across objects are established. Experimental results are given in Section 5. Finally, a conclusion is made in Section 6.

2 Previous Work

In the past, several efforts have been made for the registration of 3D point clouds. One of the most popular methods is the iterative closest point (ICP) algorithm developed by Besl and McKay [5]. The ICP searches a pair of nearest points in two data sets, and estimates a rigid transformation which aligns the two points. The rigid transformation is then applied to all the points of one data set to try to match those of the second, and the procedure is iterated until some optimisation criteria is satisfied. Several variations of the ICP method have been proposed. Chen and Medioni [7] evaluated the registration function using point-to-plane distance. In Zhang [8], a robust statistic threshold was introduced to determine the matching distance dynamically. Iterative methods such as this are obviously time consuming. When the assumption of one data set being a subset of the other is not valid, false matches can be created [9]. Moreover, they rely on a good estimate of the initial transformation. Another deficiency of the ICP method is scale sensitive. There are other alternative approaches. For example, some feature-based registration methods were presented in [10,11,12]. More detailed reviews on registration can be found in [13,14].

In face recognition, we have to register face scans of varied sizes due to either the distinct characteristics of each individual, e.g. faces between child and adult, or the scale change of a scanner. Moreover, the face surface varies significantly across subjects and often includes neck or shoulders. Finally, no transformation can be reasonably estimated to pre-align two face scans. Therefore, a non-iterative registration method addressing on those shortcomings is necessary. On the other hand, B-Spline surface fitting techniques provide potential solutions to our considerations of having compact data representation.

However, although there has been considerable work on fitting B-spline surfaces to 3D point clouds, most research is aimed at CAD or computer graphics applications, which have a different set of requirements from object recognition. Complex surfaces are often reconstructed using a network of surface patches [15,16]. Due to the uncertainty in the division of surface patches, it is difficult to establish correspondences between objects. Research on single patch surface reconstruction mostly uses structured or grid data sets with simple topology, e.g. a deformed quadrilateral [17] or a deformed cylinder [18]. The main contribution of our approach

is to have complex 3D object represented in a compact and unique way while allowing dense correspondences being established efficiently and automatically.

3 Registration

Instead of registering a probe face to a template face, our approach is to find a transformation, which takes a probe face of an arbitrary view to a canonical position (in the world coordinates system). In another word, all point clouds are in same orientation after this stage. The transformation can be written as:

$$D' = R * D = R_2 * R_1 * D \cdot \tag{1}$$

where D and D' are the point cloud before and after transformation, respectively. R is a 3×3 rotation matrix which is the composite of coarse rotation matrix R_1 and refined rotation matrix R_2 . The rotation matrix represents the pose estimate of the original data set. D' is in the canonical position where the following conditions have been satisfied:

- The line linking two inner eye corners (E_{left} , E_{right}) is perpendicular to the y - z plane after registration, Figure 1(left).
- The facial symmetry plane P is perpendicular to both the x - y plane and the x - z plane while passing through nose tip N_{tip} , nose bottom N_{bottom} and top N_{top} , Figure 1(left).
- The line linking the nose top N_{top} and nose bottom N_{bottom} is perpendicular to the x - z plane, Figure 1(right).

N_{top} is defined as the intersection of the line linking E_{left} and E_{right} and plane P .

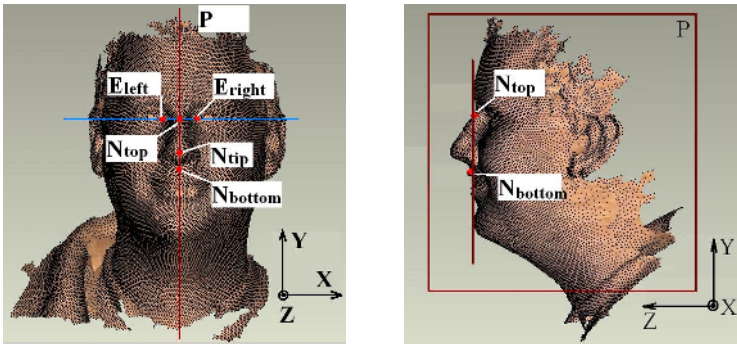


Fig. 1. Face in the canonical position after the registration

The only prior-knowledge we have before the registration stage is the location of the nose tip¹. All rest features points, i.e. inner eye corners, top and bottom point of nose, are located simultaneously with the process of pose estimation.

¹ The nose tip can be automatically located at the stage of raw data generation (via stereo matching process).

Two stages are involved to obtain the rotation matrix R and facial features. The first stage is to estimate the initial rotation matrix (head pose) based on the symmetryproperty of a face. We start from locating the plane P (Figure 2a1) which is perpendicular to both the x - y plane and the x - z plane while passing through nose tip N_{tip} . The facial profile is then extracted by the intersection of the surface and plane P , in the form of a set of scattered points, on which a B-spline curve is fitted (Figure 2b1). The candidate nose saddle point and nose bottom point can be located by calculating the first and second curve derivatives. R_{X1} is estimated by the angle between the line linking the candidate nose saddle and bottom points and the x - y plane (Figure 2b1). Figure 2c1 (side view) and Figure 2a2 (frontal view) show the result after applying the rotation matrix R_{X1} on the original data D , i.e.

$$D_1 = R_{X1} * D \tag{2}$$

Similar technique is employed to estimate R_{Y1} and R_{Z1} . Briefly, plane M in Figure 2a2 is defined as being perpendicular to both the x - y plane and the y - z plane and passing through nose tip N_{tip} . The extracted facial profile is described by a B-Spline curve on which symmetric analysis is applied. Then R_{Y1} (Figure 2b2) and R_{Z1} (Figure 2a3) are calculated. The result after applying rotation matrix R_{Y1} on D_1 is illustrated in Figure 2c2 (profile view) and Figure 2a3 (frontal view):

$$D_2 = R_{Y1} * D_1 = R_{Y1} * R_{X1} * D \tag{3}$$

The final result of stage 1 shown in Figure 2b3 (profile view) and 2c3 (frontal view) is calculated by:

$$D_3 = R_{Z1} * D_2 = R_{Z1} * R_{Y1} * R_{X1} * D = R_1 * D \tag{4}$$

Now the probe face D is near frontal after being transformed by R_1 (Figure 2c3). Next, since human faces are not perfect symmetric objects, and facial expressions also affect the symmetric measurement, the initial pose estimations need to be refined. Pose refinement uses the following rotation matrix:

$$R_2 = R_{X2} \cdot R_{Y2} \cdot R_{Z2} \tag{5}$$

where R_{X2} , R_{Y2} and R_{Z2} are the compensation rotation matrices around x , y and z axes.

The key idea of pose refinement is to evaluate R_{X2} , R_{Y2} and R_{Z2} using facial feature points. Since the coordinates of these features are directly related to pose, refining process must be done in parallel with facial features detection. With the candidate nose saddle point estimated from stage 1, possible areas containing inner corners of the eyes can then be decided upon, as shown in Figure 3a. For each area, eight candidates of the inner eye corners are obtained for further consideration (Figure 3b). The pair of points with the highest priority value is chosen as the inner eye corners (Figure 3c). The calculation of the priority is conducted under the constrains which features points must satisfy when the face is in the canonical position.

E_{left}^i and E_{right}^i denote i^{th} pair of inner eye corners from 2×8 candidates. Corresponding N_{top}^i is calculated as:

$$\theta_{Z2}^i = a \tan\left(\frac{E_{left,y}^i - E_{right,y}^i}{E_{left,x}^i - E_{right,x}^i}\right) \tag{6}$$

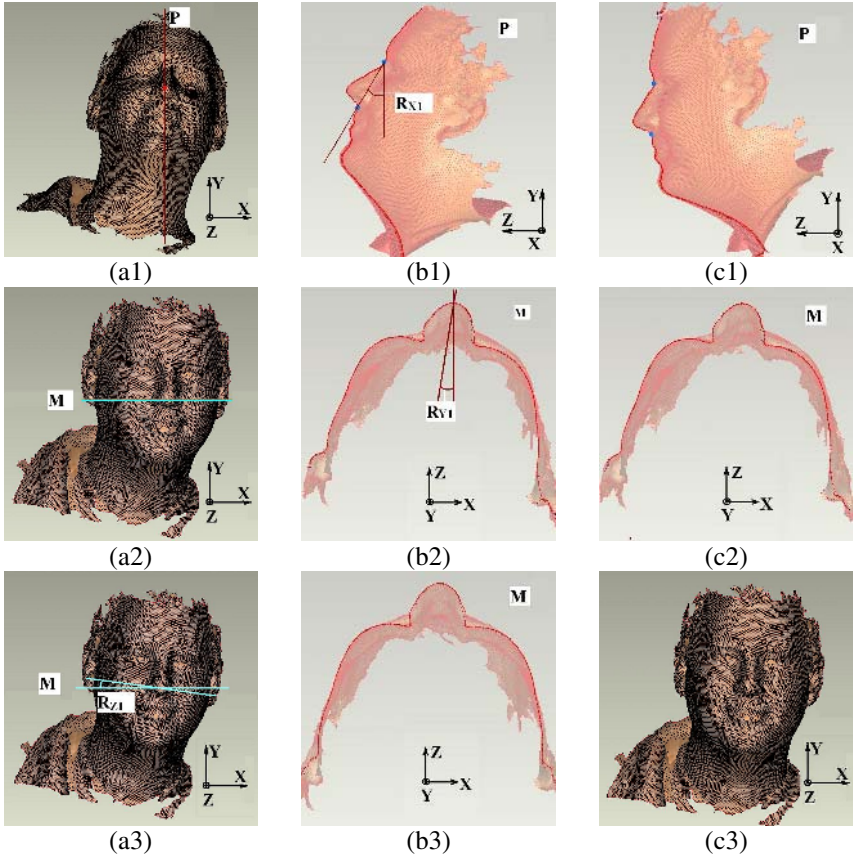


Fig. 2. Pose estimation

$$N_{top,x}^i = \frac{(E_{left,x}^i + R_{Z2}^i(\theta) \cdot E_{right,x}^i)}{2} \tag{7}$$

$$P^i = N_{top,x}^i - R_{Z2}^i(\theta) \cdot N_{tip,x} \tag{8}$$

The smaller the P^i , the higher priority the i^{th} pair has. After E_{left} and E_{right} have been decided, N_{top} is to be calculated based on the constrains of 1) having the same y -value as E_{left} and E_{right} ; 2) locating on the facial profile created by the intersection of the symmetric plane P and the surface, which is represented by B-Spline curve; 3) x -value is the mean of x -values of E_{left} and E_{right} .

$$N_{top,x}^i = \frac{(E_{left,x}^i + R_{Z2}^i(\theta) \cdot E_{right,x}^i)}{2} \tag{9}$$

$$N_{top,y} = R_y R_z E_{left,y} = R_y R_z E_{right,y} = \sum_{i=0}^m B_{i,p}(s') C_{i,y} \tag{10}$$

$$N_{top,z} = \sum_{i=0}^m B_{i,p}(s') C_{i,z} \tag{11}$$

$$N_{bottom,x} = N_{tip,x} \tag{12}$$

$$N_{bottom,y} = \sum_{i=0}^m B_{i,p}(s'')C_{i,y} \tag{13}$$

$$N_{bottom,z} = \sum_{i=0}^m B_{i,p}(s'')C_{i,z} \tag{14}$$

$$R_z N_{top,z} = R_z N_{bottom,z} \tag{15}$$

where $\sum_{i=0}^m B_{i,p}(s)C_i$ represents the face profile inferred from the facial symmetry plane.

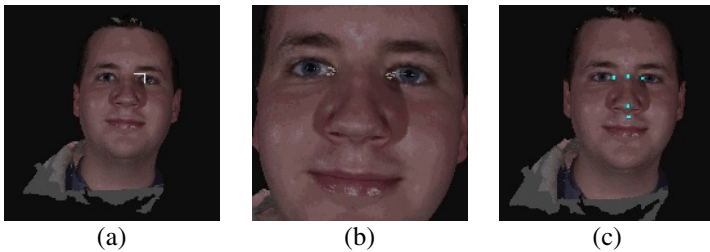


Fig. 3. Pose estimation and facial features detection. (a) Output from the first stage of pose estimation. Possible areas containing the inner corner of eyes are decided upon. (b) Candidates of the inner eye corners chosen from the areas marked in (a). (c) Detected facial features and the final output from the pose estimation algorithm.

More experimental results of comparing our 3D registration methods with the ICP algorithm are given in Section 5. Two typical examples of ICP registration are shown in Figure 4b1 and 4b2. Figure 4c1 and 4c2 are the results using our approach.

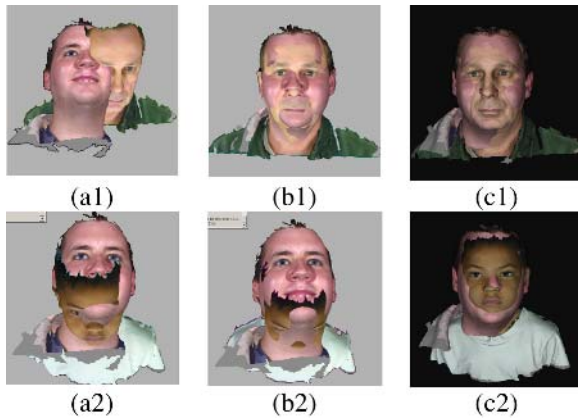


Fig. 4. Comparison between ICP method and our proposed method. (a1) First pair of point clouds to be registered. (b1) Positive result from ICP method. (c1) The registration result using our approach. (a2) Second pair of input point clouds. (b2) Negative result from ICP algorithm. (c2) Our result.

4 3D Modeling

As mentioned in the previous section, we aim to represent a complex object, e.g. a face, by a single B-Spline surface patch. This problem can be restated as follows: given an unstructured point cloud P: $p_i(x_i, y_i, z_i)$, find a single B-Spline surface Γ which fits the point cloud best. A B-Spline surface is defined as the set of points that can be obtained by evaluating the following equation for all the parameter values of s and t :

$$\Gamma(s, t) = \sum_{j=0}^n \sum_{i=0}^m B_{j,g}(s) N_{i,h}(t) C_{i,j} = p_{cd} \tag{16}$$

C is a set of control points. $B_{j,g}(s)$ is the B-Spline basis functions of degree g in the s -direction, defined over a sequence of distinguished values, known as the knot vector $U = \{u_0, u_1, \dots, u_l\}$:

$$B_{j,0}(s) = \begin{cases} 1 & \text{if } u_j \leq s < u_{j+1} \\ 0 & \text{otherwise} \end{cases} \tag{17}$$

$$B_{j,g}(s) = \frac{s - u_j}{u_{j+g} - u_j} B_{j,g-1}(s) + \frac{u_{j+g+1} - s}{u_{j+g+1} - u_{j+1}} B_{j+1,g-1}(s) \tag{18}$$

Similarly, $N_{i,h}(t)$ is defined over the knot vector $V = \{v_0, v_1, \dots, v_k\}$ in the t -direction with degree h .

4.1 Single Patch B-Spline Surface Fitting

For grid data, it is straightforward to have a B-Spline surface defined on a pair of uniform knot vectors. However, in most cases, grid data are only sufficient to represent objects with simple topology, e.g. a deformed quadrilateral or a deformed cylinder. Thus, a common approach for non-grid data representing a complex object is to divide the object surface into small patches. Each has a simple topology and can be re-gridded, on which the calculation of a pair of knot vectors is conducted and then a B-Spline patch is fitted. However, the uncertainty in the division of surface patches is the main obstacle to establish correspondences between objects.

To overcome the shortcomings above, we develop a knot vector standardisation algorithm to enable one pair of common knot vectors defined over the whole complex object surface on which a single underlying B-Spline surface can be found. Neither is the re-girding algorithm required. The knot vector standardisation algorithm is briefly illustrated by a simplified example below.

Suppose F and L are two distinctive B-Spline curves. F is defined on knot vector $X = [x_0, x_1, \dots, x_{n_x+g+1}]$ by n_x+1 control points f : $[f_1, f_2, \dots, f_{n_x}]$; L is defined on knot vector $Y = [y_0, y_1, \dots, y_{n_y+g+1}]$ by n_y+1 control points l : $[l_1, l_2, \dots, l_{n_y}]$.

$$F(x) = \sum_{i=0}^{n_x} B_{i,g}(x) f_i \tag{19}$$

$$B_{i,g}(x) = \frac{x - x_i}{x_{i+g} - x_i} B_{i,g-1}(x) + \frac{x_{i+g+1} - x}{x_{i+g+1} - x_{i+1}} B_{i+1,g-1}(x) \tag{20}$$

$$L(y) = \sum_{j=0}^{n_y} N_{j,g}(y) l_j \tag{21}$$

$$N_{j,g}(y) = \frac{y - y_j}{y_{j+g} - y_j} N_{j,g-1}(y) + \frac{y_{j+g+1} - y}{y_{j+g+1} - y_{j+1}} N_{j+1,g-1}(y) \tag{22}$$

To have $X'=Y'$ (standardised knot vectors of X and Y respectively), we standardise X and Y to a pre-defined knot vector $U=[u_0, u_1, \dots, u_{n+g+1}]$. For each element in U and X, if $x_i \in U$, x_i is untouched; If $\exists k. (u_k \in U) \cap (u_k \notin X)$, insert u_k into X; The control points f is re-calculated as $f'=[f'_1, f'_2, \dots, f'_n]^T$ and the basic function becomes:

$$B'_{k,g}(x) = \frac{x - u_k}{u_{k+g} - u_k} B'_{k,g-1}(x) + \frac{u_{k+g+1} - x}{u_{k+g+1} - u_{k+1}} B'_{k+1,g-1}(x) \tag{23}$$

The original curve is thus equal to:

$$F'(x) = \sum_{k=0}^n B'_{k,g}(x) f'_k \tag{24}$$

Similarly, control points l is re-calculated as $l'=[l'_1, l'_2, \dots, l'_n]^T$, and the basic function is re-defined on U:

$$N'_{k,g}(y) = \frac{y - u_k}{u_{k+g} - u_k} N'_{k,g-1}(y) + \frac{u_{k+g+1} - y}{u_{k+g+1} - u_{k+1}} N'_{k+1,g-1}(y) \tag{25}$$

$$L'(y) = \sum_{k=0}^n N'_{k,g}(y) l'_k \tag{26}$$

Equation 23 and 25 can be generalised in the same form:

$$Q_{k,g}(s) = \frac{s - u_k}{u_{k+g} - u_k} Q_{k,g-1}(s) + \frac{u_{k+g+1} - s}{u_{k+g+1} - u_{k+1}} Q_{k+1,g-1}(s) \tag{27}$$

Consequently, Equation 24 and 26 can be rewritten as:

$$F'(s) = \sum_{k=0}^n Q_{k,g}(s) f'_k = A(s) \bullet f' \tag{28}$$

$$L'(s) = \sum_{k=0}^n Q_{k,g}(s) l'_k = A(s) \bullet l' \tag{29}$$

where $A(s)=[Q_{0,g}(s) \ Q_{1,g}(s) \ \dots \ Q_{n,g}(s)]$. In another word, an arbitrary B-Spline curve after standardising to the common knot vector U can be represented as a vector product of A and its control points. Same technique can be applied on the pairs of knot vectors standardisation. With a common pair of knot vectors U and V defined, a single B-Spline surface Γ can be fitted on the non-grid data, which is analogously represented as:

$$\Gamma(s,t) = A(s,t) \bullet C \tag{30}$$

4.2 Correspondence

Since $A(s, t)$ is same across objects, vector C defines the unique shape of a surface, i.e. C is shape descriptors. Thus, we have established a direct mapping between the parameter domain $(s, t) \in \Omega: [0,1] \times [0,1]$ and the object space $\Gamma \in R^3$ via C . Shape descriptors have several important properties, including:

- Establishing direct one-to-one mapping from the parameter domain to the object space. For each pair of parameter value (s, t) , we have a unique corresponding B-Spline surface point in the object space.
- Affine-invariance. The same result will be obtained transforming a B-Spline surface itself or its shape descriptors.
- Shape descriptors C contain only object's geometrical properties, i.e. filtering out all the information such as location, scale and rotation attached on the object.
- Compact representation for 3D objects. The approach can achieve over 90% compression rate with similar rendering result to polygon representation. For example, the polygon representation shown in Figure 5c1 is composed of 18,649 polygons, while all the information required to rendering the smooth surface shown in Figure 5c1 is a set of shape descriptor C with the size of 616 points.

The examples of reconstructed single B-Spline representation are shown in Figure 5c1, 5c2. For comparison, we also applied the multiple B-Spline patches fitting algorithm on the same data set, Figure 5b1 and 5b2.

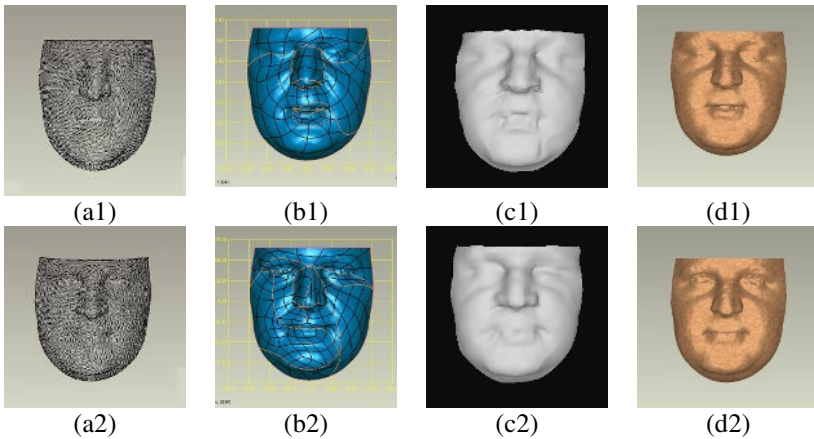


Fig. 5. Comparisons between a single B-Spline patch fitting and multiple B-Spline patches fitting. (a1) and (a2) Original data generated from different scans of the same person. (b1) Reconstructed face from (a1) composed of 140 B-Spline patches. (b2) Reconstructed face from (a2) composed of 207 B-Spline patches. (c1) and (c2) are reconstructed single B-Spline surfaces from (a1) and (a2) respectively. (d1) Polygons representation of (a1) includes 18649 polygons. (d2) Polygons representation of (a2) includes 20370 polygons.

With the one-to-one mapping from the parameter domain to the object space, the corresponding surface points between models can then be generated. Since for each pair of parameters (s, t) , each face model has a unique corresponding B-spline surface point:

$$(s, t) \Rightarrow \Gamma^k(s, t) \quad \mathbf{and} \quad (s, t) \Rightarrow \Gamma^{k+1}(s, t) \quad (31)$$

Therefore, B-Spline surface points $\Gamma^k(s, t)$ and $\Gamma^{k+1}(s, t)$ are uniquely mapped, i.e.

$$\Gamma^k(s, t) \Rightarrow \Gamma^{k+1}(s, t) \quad (32)$$

By sampling the parameter domain, e.g. uniform sampling, we obtain a set of corresponding B-Spline surface points across face models. The experimental proving is given in Section 5.

5 Experiments

5.1 3D Data Capture

All data used in our experiments were collected using a scanner based on the stereo photogrammetry technique. Currently, there are 65 subjects in our database. For each subject, 3 scans are captured. Since for a few subjects, there are fewer than 3 scans acquired, a total of 187 scans are available for conducting face recognition experiments. One in three scans of each person is used to construct the 3D face gallery whilst rest scans are used as probe face to test our face recognition system. Amongst the 65 people there are 14 females and 51 males of various ethnic background and ages. Since we were not strict on people's expression at the data collection stage, there are expressions changing from neutral to smiling among scans of individuals. The size of raw data output from the scanner is varied from 20,000 to 33,000 vertices. After the registration, the data size is reduced to average 10,000 vertices since unwanted parts, e.g. neck, shoulder etc. have been trimmed automatically.

5.2 Face Recognition Scheme

Since our face recognition experiments are conducted based on registered 3D face models, both the probe face and faces in gallery database are registered and modelled by applying the methods presented in previous sections. The procedure is briefly reviewed as follows with a gallery scan as an example. The gallery scan of unknown orientation is first automatically registered to a canonical position using the method presented in section 3. Second, the detected facial features are instinct to each subject and are taken as criterions to define the area of the frontal part of face which is then automatically separated from the unwanted part, e.g. shoulder, hair etc. Third, the modelling algorithm in section 4.1 is applied on the separated frontal part of face. Same procedures are applied to every other gallery faces, which can be done once in advance, and probe faces as well. After dense correspondences across the probe face model and the gallery models are established as proposed in section 4.2, we compare each probe face model with all the 3D face models stored in the gallery database using Euclidean distance as a matching metric. The gallery face having the smallest Euclidean distance to the probe face is identified as the best match.

With the face recognition scheme presented above, there are two main factors that may affect the recognition rate, e.g. registration error, correspondence error. We study these factors separately before we arrive at the final conclusion.

5.3 Experiments on Correspondences Establishment

The method of automatically establishing dense correspondences has been proved theoretically in section 4.2. In this section, we evaluate its errors through recognition rates. The face recognition scheme described in section 5.2, basing on the database including total 187 scans of 65 subjects is applied. However, to pinpoint the errors introduced into the face recognition system only by the corresponding method, we must rule out errors produced by other factors, e.g. registration errors, the chosen of the 65 gallery scans from 187 scans, etc. As indicated by later experimental results, constructing gallery database with scans in neutral expressions has higher recognition rates than gallery scans with other expressions. We set up the gallery database in this experiment using scans with neutral expressions. To minimise the registration errors, we manually registered every probe face to its genuine face in the gallery database in this experiment.

121 out of 122 probe faces have been correctly recognised from the gallery database of 65 subjects, i.e. the recognition rate is 99.18%. This experiment gives proofs on the correctness of our establishing correspondences approach on the practical ground.

5.4 Experiments on Registration

In this section, we conduct three experiments. While the first one focuses on testing our registration method, the other experiments using the ICP method instead are for comparison purposes. The same face recognition scheme as in Section 5.3 is adopted.

Gallery database construction: three gallery databases are available in this part of the experiments. First we construct the gallery database using scans with neutral expressions (DB1). Then the rest two of the three scans are randomly picked up to make the gallery database, DB2 and DB3 respectively.

First experiment is carried out using our non-iterative registration method in the recognition scheme. The recognition rate is 90.98% with DB1, i.e. 111 out of 122 probe faces have been correctly recognised from the gallery database of 65 subjects. Alternatively, 106 out of 122 probe faces are correctly recognised using the gallery database DB2, corresponding to the recognition rate of 86.88%; while 109 out of 122 probe faces are correctly recognised with the gallery database DB3, achieving recognition rate of 89.34%.

Two comparative experiments are conducted using ICP as the registration method in our recognition scheme. One is to register every probe face to a generic face (ICP#1), while the other is to register each probe face to its own genuine gallery face (ICP#2). The latter is merely served for a comparison purpose. To compromise that ICP method is a local method and needs a good initialisation, both processes are under careful inspection. In other words, the two ICP registrations are semi-automatic since human intervention is required to set parameters depending on individuals. Both experiments are based on the gallery database DB1. For the case of ICP#1, 59.84% recognition rate is achieved. For ICP#2, 117 out of 122 probe faces have been correctly recognised from the gallery database of 65 subjects, i.e. with recognition rate of 95.9%.

Table 1. Recognition rates under different registration methods

Registration Method	Gallery (DB1)	Gallery (DB2)	Gallery (DB3)	Status
Non-iterative	90.98%	86.88%	89.34%	Automatic
ICP#1	59.84%	---	---	Semi-automatic
ICP#2	95.9%	---	---	Semi-automatic

6 Conclusion and Future Works

We have developed a new automatic model-based face recognition system, which includes both non-iterative registration and the representation of 3D face models by shape descriptors. By registering point clouds to a canonical position, we overcome the pose-variation problem. Unlike ICP algorithm, this non-iterative registration process is scale invariant. An efficient B-spline surface-fitting technique is developed to reconstruct underlying surface for the registered data set. A new knot vector standardisation technique is proposed to allow a direct one-to-one mapping relationship from the object space to a parameter space. Subsequently, a compact parametric representation of 3D objects is obtained. The system has been tested on a personal computer (Pentium 4/512M RAM). The registration process is measured on an average sized points cloud (about 30,000 vertices), taking about 1.7 seconds. 3D modelling process takes about 0.58 seconds. Matching a probe face against 65 gallery faces can be finished within 0.02 seconds which includes the process of online correspondences establishment across the probe face and every face in the database.

Although surface distance can be used as a metric for face recognition, it may not be very sufficient since no explicit geometric information is employed. Our future work is to integrate geometric information into recognition methods. For example, with the proposed surface representation, it is possible to analyse facial component separately. As the geometry of B-spline surface can be inferred from the shape descriptors, we can delineate facial areas, e.g. forehead, nose, mouth, chin, from the parameter space, and weigh each part separately in the recognition metric to reduce the influence of facial expression. The areas potentially affected by facial expression will be given lower weight.

References

- [1] Lee, Y. and Shim, J. (2004) Curvature-based Human Face Recognition Using Depth-weighted Hausdorff Distance. International Conference on Image Processing (ICIP), pp. 1429-1432.
- [2] Lu, X., Colbry, A and Jain, K. (2004) Matching 2.5D Scans for Face Recognition. International Conference on Pattern Recognition (ICPR), pp. 362-366.
- [3] Bowyer, K., Chang, K and Flynn, P. (2006) A Survey of Approches and Challenges in 3D and Multi-modal 3D+2D Face Recognition. Computer Vision and Image Understanding, 101, 1-15.
- [4] Campbell, R. and Flynn, P. (2001) A Survey of Free-form Object Representation and Recognition Techniques. Computer Vision and Image Understanding, vol. 81, pp. 166-210.

- [5] Besl, P.J., and McKay, N.D. (1992) A Method for Registration of 3D Shapes. *IEEE Pattern Analysis and Machine Intelligence*, vol. 14, No. 2, 239-256.
- [6] Medioni, G. and Waupotitsch, R. (2003) Face Recognition and Modelling in 3D. *IEEE International Workshop on Analysis and Modelling of Faces and Gestures (AMFG)*, pp. 232-233.
- [7] Chen, Y., and Medioni, G. (1992) Object modelling by registration of multiple range images. *Image and Vision Computing*, vol. 10, No. 3, 145-155.
- [8] Zhang, Z. (1994) Iterative Point Matching for Registration of Free-form Curves and Surfaces. *International Journal of Computer Vision*, vol. 13, No. 2, pp.119-152.
- [9] Fusiello, A., Castellani, U., Ronchetti, L., and Murino, V. (2002) Model Acquisition by Registration of Multiple Acoustic Range Views, *Computer Vision, ECCV2002*, Springer, pp. 805-819.
- [10] Godin, G., Rioux, M. and Baribeau, R. (1994) Three-dimensional Registration Using Range and Intensity Information, *SPIE*, vol. 2350, *Videometrics III*, pp. 279-290.
- [11] Godin, G. and Boulanger, P. (1995) Range Image Registration Through Viewpoint Invariant Computation of Curvature, *IAPRS*, 30 (5/W1), pp. 170-175.
- [12] Godin, G., Laurendeau, D. and Bergevin, R. (2001) A Method for the Registration of Attributed Range images, *International Conference on 3D Imaging and Modeling*, Quebec, pp. 179-186.
- [13] Campbell, R., Flynn, P. (2001) A Survey of Free-form Object Representation and Recognition Techniques, *Computer Vision and Image Understanding*, vol. 81, pp. 166-210.
- [14] Flusser, J. and Zitova, B. (2003) Image Registration Methods: A Survey, *Image and Vision Computing*, vol. 21, pp. 977-1000.
- [15] Eck, M., and Hoppe, H. (1996) Automatic Reconstruction of B-Spline Surfaces of Arbitrary Topological Type. *Proc. 23rd Int'l. Conf. on Computer Graphics and Interactive Techniques SIGGRAPH '96*, ACM, New York, NY. pp. 325-334.
- [16] Krishnamurthy, V. and Levoy, M. (1996) Fitting Smooth Surfaces to Dense Polygon Meshes, *ACM-0-89791-746-4/96/008*.
- [17] Sarkar, B. and Menq, C. (1991) Parameter Optimization in Approximating Curves and Surfaces to Measurement Data, *Computer Aided Geometric Design*, vol. 8, pp. 267-290.
- [18] Forsey, D. and Bartels, R. (1995) Surface Fitting with Hierarchical splines, *ACM Transactions on Graphics*, vol. 14, no. 2, pp. 134-161.

Automatic Denoising of 2D Color Face Images Using Recursive PCA Reconstruction

Hyun Park and Young Shik Moon

Department of Computer Science and Engineering, Hanyang University,
1271 Sa-Dong, Ansan, Kyunggi-Do 425-791, Korea
{hpark, ysmoon}@cse.hanyang.ac.kr

Abstract. In this paper, we propose a denoising method based on PCA reconstruction for removing complex color noise components on human faces, which is not easy to remove by using vectorial color filters. The proposed method is composed of the following six steps: training of canonical eigenface space using PCA, automatic extraction of facial features using active appearance model and alignment of the input face to mean shape, reconstruction of an initial noise free face, relighting of reconstructed face using a bilateral filter, extraction of noise regions using the variances of skin color of training data, and reconstruction using partial information of input images (except the noise regions) and blending of the reconstructed image with the original image. Experimental results show that the proposed denoising method maintains the structural characteristics of input faces, while efficiently removing noise components with complex colors.

1 Introduction

Denoising and reconstruction of color images have been extensively studied in the field of computer vision and image processing. There have been some attempts to remove noises on color images. Early attempts removed noises on color images through independent smoothing of RGB channels. Generally, almost all approaches focus on a variety of filtering processes applied appropriately to the color vectorial data. The color filters such as vector median and directional filters are used for removing Gaussian white noise or impulse noise in the field of computer vision and image processing [1], [2], [3]. The nonlinear color filters such as WMF (weighted median filter) and CWMF (center weighted median filter) efficiently remove impulse and salt & pepper noises [4], [5]. The decomposition using PCA (principal component analysis), kernel PCA, ICA (independent component analysis) and wavelet transform is also applied to denoising [6], [7], [8]. In spite of these efforts, many of denoising methods have been performed on gray images and they removed mostly simple noises such as gaussian noises or impulse noises. Moreover, complex color noise components on human faces are difficult to remove by general color filtering processes.

Therefore, we propose a new denoising method based on recursive PCA reconstruction, which maintains the structural characteristics of input face and efficiently removes complex color noise components on input faces. The proposed method is

composed of the following four steps. First, we construct a canonical eigenface space using PCA. Next, we automatically extract facial features of an input face using the multi-level active appearance model (MAAM). To minimize the reconstruction error by geometric misalignment, we align the input face to the reference shape using the extracted facial feature points. Next, we reconstruct an initial noise free face by projecting the input face onto constructed canonical eigenface space. We carry out the proposed relighting method so that both the reconstructed face and the input face have the same illumination condition. Then, we extract noise regions using the variances of vector magnitude and vector angle of the skin color at the each pixel position of the training data. Finally, if the extracted noise regions are less than 40% of the total face region, we reconstruct the noise free face once again using partial information of the input image (except the noise regions), and the reconstructed noise free face is blended appropriately with the original image.

2 Training of Canonical Eigenface Space Using PCA

Generally, the original eigenface space is effective for recognition and reconstruction, but it is not robust against various illumination changes and geometric misalignments. In this paper, complex color noise components on training faces are manually removed, then training images are aligned to the mean shape of AAM and normalized to ‘zero mean and unit length’, as in equation (1).

We refer to this training face as the normalized canonical face. The training data consists of still images of 100 different frontal-view human faces, all without glasses and with a neutral expression. We construct a canonical eigenface space using PCA and normalized canonical training faces. Finally, the canonical eigenface space is constructed by removing the rest of complex color noises that have not been removed manually, which is performed by selecting only 95% of the principal components.

$$x = \frac{X - E\{X\}}{\|X - E\{X\}\|}, \quad E\{X\} = \bar{x} = \frac{1}{N} \sum_{n=1}^N x_n, \quad \Sigma = \frac{1}{N} \sum_{n=1}^N (x_n - \bar{x})(x_n - \bar{x})^T \quad (1)$$

where x is the normalized face of original face X , \bar{x} is the mean vector, Σ is the covariance matrix, N is the number of faces in the training set.

$$\Lambda = \text{diag}(\sigma_i^2) = \Phi^T \Sigma \Phi, \quad \alpha = \Phi_m^T (x - \bar{x}), \quad x^* = \sum_{i=1}^m \alpha_i \phi_i \quad (2)$$

In equation (2), Λ is a diagonal matrix in which diagonal terms are eigenvalues of Σ , and σ_i^2 is the variance of training faces in the direction of i th eigenvector. Φ is an eigenvector matrix, α is a principal component vector, m is the number of eigenvectors, and x^* is the reconstructed face by projection onto the canonical eigenface space [6], [9]. If we use partial information of the input face and the scaled eigenvectors $\sigma_i \phi_i$ as a basis, reconstructed face x^* is defined by equation (3) [9].

$$x^* = \sum_{i=1}^m \alpha_i \sigma_i \phi_i = \Phi \cdot \text{diag}(\sigma_i) \cdot \alpha \quad (3)$$

3 Automatic Extraction of Facial Feature Points Using AAM

In order to improve the efficiency and robustness of the matching algorithm, a facial feature template is matched by using the MAAM (Multi-level active appearance model) based on color images. The training method for AAM uses the Jacobian learning scheme. As in the original AAM method, these AAMs are built at each level of a scale-pyramid for coarse-to-fine fitting based on multi-resolution [10], [11].



Fig. 1. Face textures and shapes (landmarks) for training the AAM

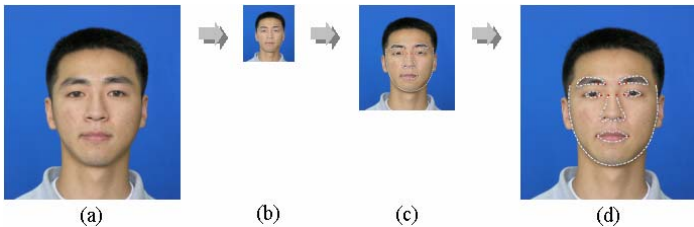


Fig. 2. Search-based initialization of deformation templates using multi-level AAM. (a) Original image. (b) Search result in level 2 (50 % scaled down of level 1). (c) Search result in level 1 (50 % scaled down of level 0). (d) Final search result in level 0 (not scaled down).

The training faces are acquired in 1187 x 1190 bitmap color format. As shown in Figure 1(b), the facial structures are manually annotated using 94 total landmarks of eyebrows, eyes, nose, mouth, and jaw. As shown in Figure 2, a multi-resolution pyramid is scaled into three levels. Figure 2(d) is the final search result in which parameters of the combined model for AAM are optimized (translation, scaling, rotation, texture model parameters, and shape model parameters).

4 Reconstruction of Noise Free Face Reflecting Various Facial Colors

In the PCA reconstruction in color domain, the difference between the color distribution of input face and that of training faces causes PCA reconstruction error. Especially, the reconstruction error becomes larger as the illumination condition changes. Therefore, the direct projection of color face onto the canonical eigenface space, using x^* of equation (2), may not be robust. We solve this problem by using the polynomial regression approximation.

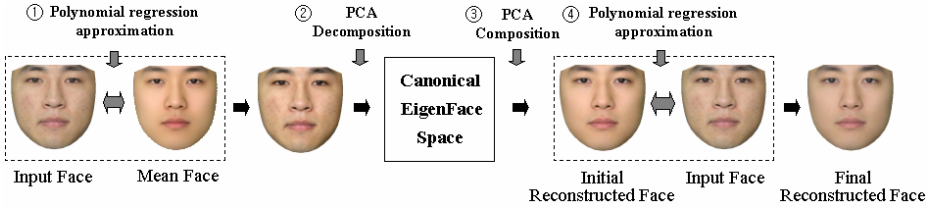


Fig. 3. PCA reconstruction reflecting various facial colors

As shown in ① of Figure 3, in order to minimize the difference of facial color between the input face and the training faces, we convert the facial color of input face to the facial color of the mean face using the Polynomial-Least Squares Fitting (PLSF) that is carried out by equation (5) and (6). Next, by projecting the input face onto the canonical eigenface space, the initial noise free face is reconstructed, as in ②, ③ of Figure 3. Finally, we convert the facial color of reconstructed face to the facial color of the input face.

$$\begin{aligned}
 \mathbf{x}_{facial\ color\ A} &= \{x_1, x_2, x_3, \dots, x_n\} \\
 \mathbf{y}_{facial\ color\ B} &= \{y_1, y_2, y_3, \dots, y_n\} \\
 \mathbf{y}_{facial\ color\ A} &= \{y'_1, y'_2, y'_3, \dots, y'_n\}
 \end{aligned} \tag{4}$$

$$\mathbf{y}_{facial\ color\ A} = PLSF(\mathbf{x}_{facial\ color\ A}, \mathbf{y}_{facial\ color\ B}) = ax^2 + bx + c = y' \tag{5}$$

$$a = \frac{(n \sum x^2 y) - (\sum x^2 \sum x)}{(n \sum x^4) - (\sum x^2)^2}, b = \frac{\sum xy}{\sum x^2}, c = \frac{\sum x^4 \sum y - \sum x^2 \sum x^2 y}{(n \sum x^4) - (\sum x^2)^2} \tag{6}$$

where $\mathbf{x}_{facial\ color\ A}$ is a reference face with facial color A, $\mathbf{y}_{facial\ color\ B}$ is an input face with facial color B, $\mathbf{y}_{facial\ color\ A}$ is a converted input face.

5 Relighting of Reconstructed Face

We can not always assume that both the reconstructed face and the input face are on the same illumination condition. To compensate the difference of illumination conditions, we propose a relighting method using the bilateral filter in HSV color domain [12]. Our relighting method is composed of the following two steps: initial relighting of reconstructed face using the bilateral filter, compensation of the initial relighting to exclude noise influences.

5.1 Relighting of Reconstructed Face Using Bilateral Filter

The bilateral filter combines a classic low-pass filter with an edge-stopping function that attenuates the filter kernel weights when the intensity difference between pixels is

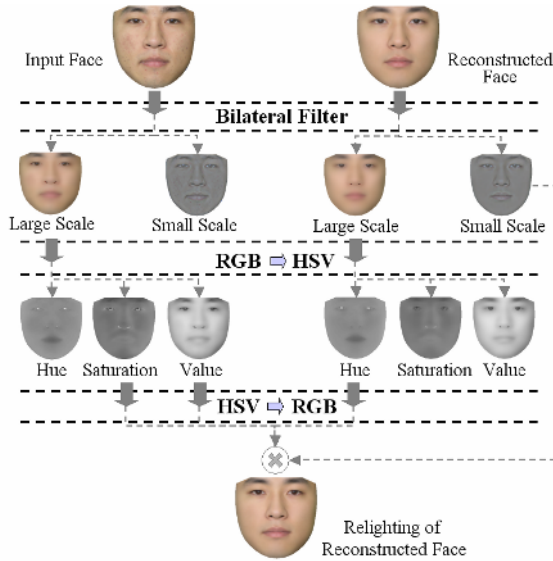


Fig. 4. Relighting of reconstructed face image

large. As shown in Figure 4, the bilateral filter separates a color face image into the small scale and the large scale. The small scale represents the detailed shape of image and the large scale represents the illumination of image. We apply the bilateral filter to each RGB color channel separately with the same standard deviation parameters for all three channels. The output of bilateral filter using equation (7) and (8) is the large scale. The small scale is computed to divide color image by its large scale. By combining the saturation and value (brightness) of the large scale input face, the hue of large scale reconstructed face, and the small scale reconstructed face, we perform the relighting of the reconstructed face. The computation is done in the log domain to take the intensity ratios in account. Spatial variance σ_f is equal to 2% of the image diagonal and variance σ_g is equal to 0.4 for intensity influence [13], [14].

$$J_s = \frac{1}{k(s)} \sum_{p \in \Omega} f(p-s) g(I_p - I_s) I_p \quad (7)$$

$$k(s) = \sum_{p \in \Omega} f(p-s) g(I_p - I_s), \quad f(x) = g(x) = \frac{1}{2} e^{-\left(\frac{x^2}{\sigma^2}\right)} \quad (8)$$

Here p and s are pixel positions, I_p and I_s are pixel values (or each color channel values) at p and s pixel positions respectively, $k(s)$ is normalization term and Ω is the size of filter mask.

5.2 Compensation of Relighting Using Extracted Noise Region Information

The large scale representing the illumination of color face is affected by complex color noise components that are difficult to remove by Gaussian filtering of the

bilateral filter. Therefore, the result of relighting might have a little noise effects. To prevent these noise effects occurred by the initial relighting, we propose the modified joint bilateral filter using noise region such as equation (9), (10) and (11).

$$J_s = \frac{1}{k(s)} \sum_{p \in \Omega} f(p-s)g(e(I_p)-e(I_s))e(I_p) \tag{9}$$

$$k(s) = \sum_{p \in \Omega} f(p-s)g(e(I_p)-e(I_s)) \tag{10}$$

$$e(I_p) = \begin{cases} \text{if } p \in \text{noise region, the pixel of } p \text{ index on reconstructed face} \\ \text{if } p \notin \text{noise region, the pixel of } p \text{ index on input face} \end{cases} \tag{11}$$

6 Extraction of Noise Regions Using the Variance of Skin Color

In order to extract complex color noise regions automatically, we propose a noise detection method based on the vector magnitude map (VMM) and vector direction map (VDM) of training data that is computed by equation (12) and (13) [2], [3]. The vector magnitude represents the distance and the vector direction represents the angle between two vectors at the same pixel position. As in equation (14), we use the standard deviation of VMM and VDM at each pixel position as the threshold value. The threshold value for noise extraction is determined by experiments for over detection of color noise regions. So the threshold value has been empirically selected to 2 times standard deviation of VMM and VDM, based on noise detection rate.

$$\sigma_{VM\ i}^2 = \frac{1}{N} \sum_{n=1}^N (D(x_{n,i}, \bar{x}_i) - \overline{D(x_{n,i}, \bar{x}_i)})^2, \quad D(a,b) = \left(\sum_{k=1}^d (a^k - b^k)^2 \right)^{1/2} \tag{12}$$

$$\sigma_{VD\ i}^2 = \frac{1}{N} \sum_{n=1}^N (A(x_{n,i}, \bar{x}_i) - \overline{A(x_{n,i}, \bar{x}_i)})^2, \quad A(a,b) = \cos^{-1} \left(\frac{a \cdot b^T}{|a||b|} \right) \tag{13}$$

$$x_{input,i} \text{ is noise pixel if } \begin{cases} 2\sigma_{VM\ i} < D(x_{input,i}, x_{recon,i}) \text{ and} \\ 2\sigma_{VD\ i} < A(x_{input,i}, x_{recon,i}) \end{cases} \tag{14}$$

Here \bar{x} is the mean vector of the training set, N is the number of data in the training set, $D(a,b)$ is the distance between two vectors, $A(a,b)$ is the angle between two vectors, a and b are pixel vectors, k is the dimension of a pixel vector, $\sigma_{VM\ i}$ is the standard deviation of VMM at i th position and $\sigma_{VD\ i}$ is the standard deviation of VDM at i th position.

7 Reconstruction Using Partial Information and Blending

Generally, the least squares minimization (LSM) method using orthogonal projection and the original PCA reconstruction are not robust when input images have

intra-sample outliers. We can regard complex color noise components on facial images as intra-sample outliers. Therefore, we reconstruct the optimal noise free face using the robust PCA based on the singular value decomposition (SVD). The robust PCA computes the optimal principal components by using the partial information of input images (except the noise regions), and then we construct the noise free face by using the optimal principal components [9], [15]. If the area of the extracted noise region is more than 40% of the total face region, we use the reconstructed noise free face by orthogonal projection.

We define an error function $E(\alpha)$ such as equation (16), as the sum of square errors which are the difference between pixel values in non-noise regions and its reconstructed ones. Our goal is to find the optimal α^* so as to minimize the error.

$$\alpha^* = \arg \min_{\alpha} E(\alpha) \tag{15}$$

$$E(\alpha) = \sum_{j=1}^p \left(\tilde{x}(j) - \sum_{i=1}^m \alpha_i \sigma_i \phi_i(j) \right)^2 \tag{16}$$

Here $\tilde{x}(j)$ are pixels of the input image except the noise regions, p is the number of pixels in the non-noise regions. If $q_i = \sigma_i \phi_i$, the equation (16) is replaced by equation (17).

$$E(\alpha) = \sum_{j=1}^p \left(\tilde{x}(j) - \sum_{i=1}^m \alpha_i q_i(j) \right)^2 = |\tilde{x} - Q\alpha|^2 \tag{17}$$

$$Q = UWV^T, \quad Q^+ = VW^+U^T \tag{18}$$

$$W^+ = \text{diag} \begin{pmatrix} w_i^{-1} & \text{if } w_i \neq 0 \\ 0 & \text{otherwise} \end{pmatrix} \tag{19}$$

$$\alpha = Q^+ \tilde{x}, \quad \alpha^* = \sigma \alpha = \sigma Q^+ \tilde{x} \tag{20}$$

$$x^{recon} = \sum_{i=1}^m \alpha_i^* \phi_i + \bar{x} \tag{21}$$

We can get the optimal α^* by using the pseudo-inverse of Q that is computed by SVD. As the reconstructed face by orthogonal projection, we apply the facial color transfer and the relighting to the reconstructed face by equation (21). Finally, we blend the reconstructed noise free face with the original face by equation (22). In equation (22), b is a blending ratio. In this paper, we use 0.9 as the blending ratio. O is the original face, and R is the final reconstructed face.

$$\text{Final noise free face} = \begin{cases} b \cdot O + (1-b) \cdot R & \text{if noise free region} \\ (1-b) \cdot O + b \cdot R & \text{if noise region} \end{cases} \tag{22}$$

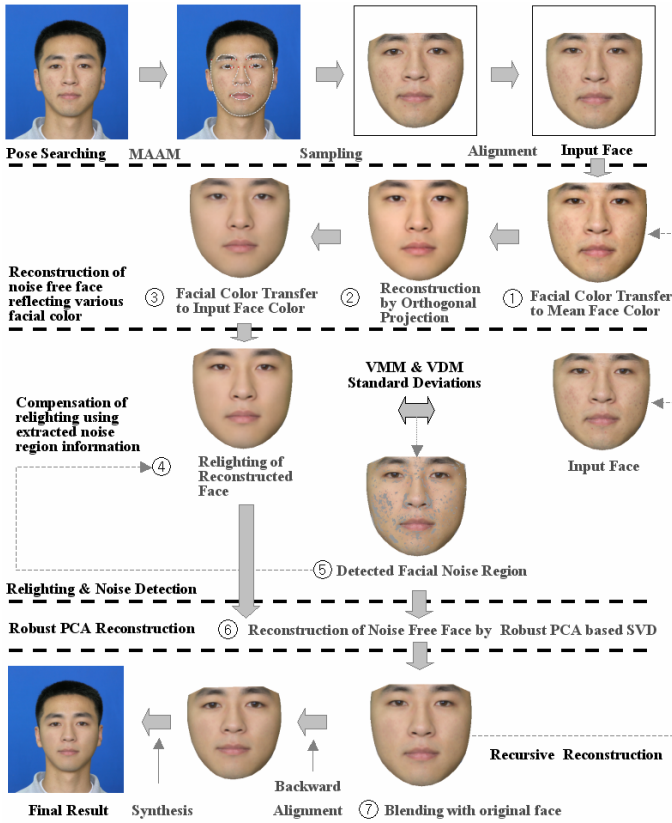


Fig. 5. Denoising process based on recursive PCA reconstruction

8 Experimental Results

We evaluate the performance of the proposed denoising method by removing noise components on the frontal face. For this experiment, we manually insert complex color noise components such as pockmarks, pimples and blotches on a clear face. As shown in Figure 6, the proposed denoising method maintains the structural characteristics of input face, while efficiently removing complex color noise components. As we carry out the denoising process repeatedly, more detailed information on the face is blurred. However, this blurring is negligible. As shown in Figure 7, we also evaluate the performance of the proposed denoising method by comparing with multilevel inpainting method, TV inpainting method, 7x7 WMF and 7x7 CWMF. Experimental results show that the proposed denoising method is more efficient in terms of smoothness, visual impression and denoising effect than the other methods.



Fig. 6. Noise removal results by using recursive PCA reconstruction (a) noise free input face, (b) input face with arbitrary noise components, (c) reconstructed results when the removal is performed once, (d) reconstructed results when the removal is performed three times

9 Conclusions

In this paper, we propose a denoising method based on PCA reconstruction for removing complex color noises on human faces, which is difficult to remove by using general color filters. The proposed method maintains the structural characteristics of input faces, while efficiently removing complex color noises on input faces. Experimental results show that the proposed denoising method efficiently removes complex color noise components on input face images.



Fig. 7. Comparison of original face with noise removed results by proposed method, inpainting methods and spatial filter methods (a) noise free input face, (b) input face with arbitrary noise components, (c) result by the proposed method (executed once), (d) result by the proposed method (iterated three times), (e) result by multilevel inpainting method, (f) result by TV inpainting method, (g) result by 7x7 WMF, (h) result by 7x7 CWMF

Acknowledgement

This research was supported by the MIC (Ministry of Information and Communication) of Korea, under the ITRC (Information Technology Research Center) support program supervised by the IITA (Institute of Information Technology Assessment).

References

1. Ben-Shahar, O., Zucker, S. W.: The Perceptual Organization of Texture Flow: A Contextual Inference Approach. *IEEE Trans. on PAMI*, Vol. 25, No. 4, (2003) 401-417
2. Lukac, R., Smolka, B., Martin, K., Plataniotis, K. N., Venetsanopoulos, A. N.: Vector Filtering for Color Imaging. *IEEE Signal Processing Magazine*, Special Issue on Color Image processing, Vol. 22, No. 1, Jan, (2005) 74-86
3. Lukac, R., Plataniotis, K. N.: A Taxonomy of Color Image Filtering and Enhancement Solutions, In *Advances in Image and Electron Physics*, Vol. 140, Feb/Mar, (2006) 187-264
4. Arce, G. R.: *Nonlinear Signal Processing : A Statistical Approach*, WILEY, (2005)
5. Bovik, A.: *Handbook of Image & Video Processing*, Elsevier Academic Press, (2005) 109-127
6. Li, X., Ning, Z., Xiang, L.: Robust 3D Reconstruction with Outliers Using RANSAC Based Singular Value Decomposition. *IEICE Trans. on Information and Systems*, (2005)
7. Takahashi, T., Kurita, T.: Robust De-nosing by Kernel PCA. *ICANN 2002*, LNCS 2415, (2002) 739-744
8. Hyvärinen, A., Hoyer, P., Oja, E.: Sparse Code Shrinkage for Image Denoising, *IEEE Int'l Joint Conf. on Neural Networks Proceedings*, Vol. 2, May, (2001) 859-864
9. Blanz, V., Mehl, A., Vetter, T., Seidel, H. P.: A Statistical Method for Robust 3D Surface Reconstruction from Sparse Data. *IEEE 3DPVT'04*, (2004) 239-300
10. Stegmann, M. B. B., Ersboll, K., Larsen, R.: FAME-A Flexible Appearance Modelling Environment, *IEEE Trans. on Medical Imaging*, Vol. 22, (2003) 1319-1331
11. Cootes, T. F., Taylor, C. J.: *Statistical Models of Appearance for Computer Vision*, Tech. Report, University of Manchester, <http://www.isbe.man.ac.uk/~bim/>, Feb, (2000)
12. Wyszecki, G., Stiles, W. S.: *Color Science, Concepts and Methods, Quantitative Data and Formulas*, John Wiley, N. Y., 2nd Edition, (1982)
13. Eisemann, E., Durand, F.: Flash Photography Enhancement via Intrinsic Relighting, *ACM SIGGRAPH*, Vol. 23, (2004) 673-678
14. Petschnigg, G., Agrawala, M., Hoppe, H., Szeliski, R., Cohen, M., Toyama, K.: Digital Photography with Flash and No-Flash Image Pairs. *ACM SIGGRAPH*, (2004) 664-672
15. Hwang, B. W., Lee, S. W.: Reconstruction of Partially Damaged Faces Based on a Morphable Face Model. *IEEE Trans. on PAMI*, Vol. 25, No. 3, (2003) 365-372

Facial Analysis and Synthesis Scheme

Ilse Ravyse and Hichem Sahli

Vrije Universiteit Brussel, Department ETRO,
Audio Visual Signal Processing (AVSP)
Pleinlaan 2, 1050 Brussel
{icravyse, hsahli}@etro.vub.ac.be
<http://www.etro.vub.ac.be>

Abstract. We developed an algorithmic scheme to extract the semantical description of the face and the face motion from an image sequence, and to re-play this action in a 3-dimensional (3D) virtual world. The presented *Facial Analysis and Synthesis Scheme* combines new methods for detection and tracking of the face and facial features, for estimating the 3D face movements and the nonrigid facial expressions, and for extracting the MPEG4 facial animation parameters. In the scheme, the face is treated either as a 2D object that has specific color, shape and motion characteristics, either as a 3D model that is calibrated and moved using a natural displacement-based deformation model. A dynamic MPEG4 displacement table takes care of the semantical controls of the animations of the face model. As a result, this virtual face model mimics well the gestures of the person in the video.

1 Introduction

The communication of the non-verbal face gestures is used in a wide range of applications as tele-presence and surveillance, and as an element in the upcoming new media in games and intuitive user-interfaces with virtual actors. The corresponding digital content contains a large amount of data, namely natural recordings and/or synthetic data, which have to be stored or sent efficiently. This paper addresses analysis techniques which allow to replace the raw video recording of a person by a 'high level' semantical representation and to drive a face model according to the face appearance.

1.1 The Problem Formulation

The goal of the *Facial Analysis and Synthesis Scheme* (FASS) is to estimate the static and dynamical parameters that respectively correspond to the structure and motion of the face. What makes this problem challenging is that 2D and 3D information have to be extracted from a single 2D image sequence. Both the rigid motion and the expressions of the face in the image sequence are extracted to control a virtual face model. To this end, several restrictions are imposed on the scheme: only natural, mechanical motion models can be used which allow to

explain the face motions; a physical face model based upon biological evidence has to be used; the final face motion representation needs to be conform to the MPEG4 video compression standard [1,2]; and the algorithms have to work as automatic as possible. Therefore the 3D model animation is preferred as framework over the image-based performance-based animation techniques [3].

The remainder of this paper is organized as follows. In section 2 all the building blocks of the FASS are discussed, and in section 3 conclusions are drawn.

2 Scheme

An entire scheme is proposed to study the face motions recorded in a (color) image sequence and to render them in a virtual world. Its building blocks, given in Figure 1, combine techniques from computer vision, computer graphics and mechanics as explained in the following sections.

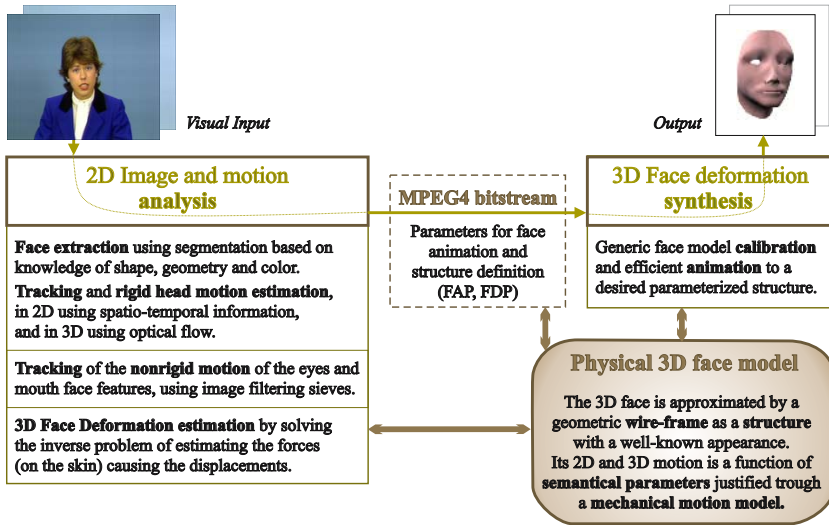


Fig. 1. Functional blocks of the FASS

2.1 Face Extraction

The localization of the face in the first image is performed by an automatic segmentation and verification using both the face color and spatial characteristics [4]. A pixel is initially labeled as skin if its color falls inside the boundary of the delimited skin color region in the YC_bC_r space. Because of each person's individual skin color, that initial segmentation is personalized by selecting all regions of skin segments with high skin probability as face candidates. This probability is the value of a C_bC_r Gaussian fitted on the initial detected skin chromaticity and skewed towards reddish (small C_b , large C_r), as depicted in Figure 2a.

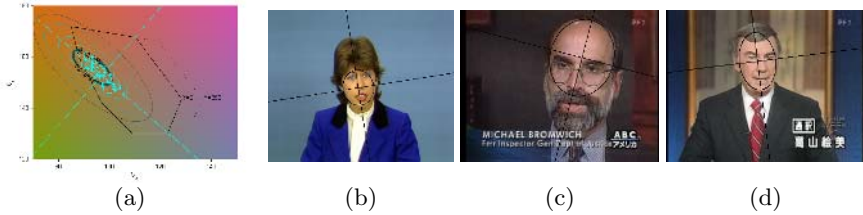


Fig. 2. (a) Skin color region in the $YCbCr$ space, where the highest skin probabilities are found inside the threshold ellipse of the skewed Gaussian fit; (b-d) Localized faces

A global face cue measure is then estimated for each candidate region. It is the sum of the z-scores of the shape cues, derived from an ellipse fit on each region, and the gray-tone cues that express smoothness and the existence of facial feature corners in a face region. The face candidate that has the maximal measure localizes the face in the image. The construction of that measure allows it to be easily extendible with other cues and to be adapted towards the face in the input image. Experiments on a database (of [5]) resulted in 92% of good face detections, which is quite powerful compared to the 93% of the state-of-the-art face detection (using appearance training) of Viola [6]. Some examples of our face extraction algorithm are shown in Figure 2b-d

2.2 2D Head Tracking

The tracking of the detected head in the subsequent image frames is performed via a kernel-based method wherein a joint spatial-color probability density characterizes the ellipse head region [4]. The parameterized motion and the illumination changes affecting the target are estimated by minimizing a distance measuring the adherence of the head candidate to the density of the head model. This kernel-based approach proved to be robust to the 3-dimensional motion of the face, and lets the tracked region remain tightly around the face as shown in Figure 3. Moreover, incorporating an illumination model into the tracking equations enables us to cope with potentially distracting illumination changes. The proposed algorithm achieves reliable tracking results compared to the best spatially-weighted color histogram trackers [7,4]. The robustness of the joint spatial-color tracker against a background of the same histogram is illustrated in Figure 4.

2.3 Facial Feature Motion

Facial features play a special role in recognizing a specific face, but also in following the motion of the face. Although it is hard to extract individual face feature points, the shape of the eye can be analyzed in local windows of the intensity images obtained by an eye detection algorithm of template matching. By using morphological scale-space, information about edges and regions of the eye is extracted from these images. Starting from the approach of Matthews [8], we



Fig. 3. Joint spatial color tracker on a face image sequence

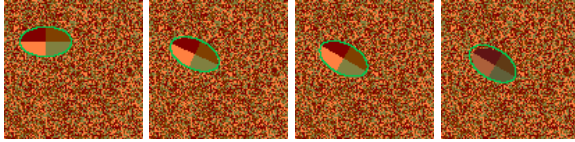


Fig. 4. Joint spatial color tracker of a target on a background with same histogram, changing its illumination in the fourth frame

proposed modifications to incorporate additional information about the spatio-temporal nature of an eye gesture [9]. Namely, for the estimation of the eye opening/closing animation parameters, we introduced a gesture observation measure that depends both on the scale of the eye region details and on the eye blink timing. In a typical newsreader sequence, we can automatically detect the eyes regions and estimate eyes gestures states, as shown in Figure 5. This algorithm also allows estimating the mouth opening of the person in the face image sequence.

2.4 Calibration in 3D

A 3D wire-frame face model provides depth and topology information that aids the 3D face analysis. In our work we use a generic face model that has a topology for handling face motion, and the calibration ensures that this model has the looks and 3D position of the face in the first frame of the image sequence. While Pighin et al. [10] aimed at photorealistic face reconstruction, using a dense laser-scanned face model and multiple views of a person, we employ the calibration as a preprocessing step for the natural motion estimation of section 2.5.

Based upon the semantical correspondence between the face feature's pixels and vertices, the camera and initial positioning parameters are assigned, as well as the initial structure deformation. All visible model vertices $\{\mathbf{X}_n = (x_n, y_n, z_n)\}_{n=1}^N$ are mapped on the image pixels $\{\mathbf{x}_n = (i_n, j_n)\}_{n=1}^N$ by a perspective projection with fixed focal length f . The projected face is correctly placed on the image face region by changing the image center of projection \mathbf{c} , while its orientation and scale are provided by a rigid motion of the face model which consists of a small rotation with vector $(0, 0, \omega_z)$ and a depth translation $(0, 0, t_z)$ relative to the axes of projection. These parameters are estimated such that M feature correspondences (with $M < N$) are fulfilled in least-squares sense:

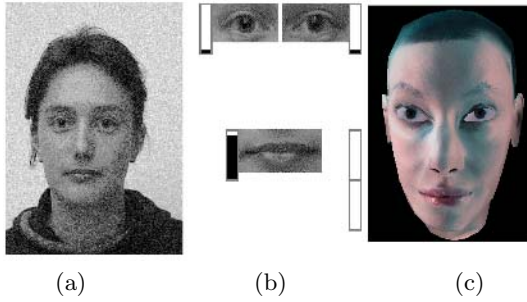


Fig. 5. Face feature gesture analysis: (a) the newsreader, (b) the eye and mouth windows with gesture state bars (fully black means a closed feature), and face tilt angle in the double bar, (c) the mimic of the virtual model

$$(\widehat{\mathbf{c}}, \widehat{\omega}_z, \widehat{t}_z) = \underset{(\mathbf{c}, \omega_z, t_z)}{\arg \min} \sum_{m=1}^M \left\| \begin{pmatrix} i_m \\ j_m \end{pmatrix} - \frac{f}{z_m + t_z} \begin{pmatrix} x_m - \omega_z y_m \\ -(\omega_z x_m + y_m) \end{pmatrix} + \mathbf{c} \right\|^2 \quad (1)$$

Volume morphing adjusts the face model’s geometrical structure to the appearance of the frontal looking person in the image. To create a face model of that person, a tri-variate 3D radial basis function (r.b.f.) interpolation [11] displaces the vertices which are not given by the feature correspondence. Depth information is either copied from the original 3D model, or selected from a side view of the person. A sensitivity study of the influence of the new feature point position on all calibrated face model vertices has shown that the linear r.b.f. acts upon a local region around the feature point, lets the face surface remain smooth after morphing and does not distort the face model topology. The displaced model vertices $\{\mathbf{Y}_n\}_{n=1}^N$ and its given feature point subset $\{\mathbf{L}_m\}_{m=1}^k$ are described in terms of their original positions, given by respectively $\{\mathbf{X}_n\}_{n=1}^N$ and $\{\mathbf{K}_m\}_{m=1}^k$, as

$$\mathbf{Y} = (\mathbf{L}_1 \dots \mathbf{L}_k) \left(\mathbf{A}^T \begin{pmatrix} \frac{\|\mathbf{X} - \mathbf{K}_1\|_2}{\sigma_1} \\ \vdots \\ \frac{\|\mathbf{X} - \mathbf{K}_k\|_2}{\sigma_k} \end{pmatrix} + \mathbf{B}^T \right) \quad (2)$$

where \mathbf{A} is a $k \times k$ and \mathbf{B} is a $1 \times k$ matrix constructed from the symmetric r.b.f. kernel matrix $\mathbf{H} = \{H_{ml} \triangleq \frac{\|\mathbf{K}_m - \mathbf{K}_l\|_2}{\sigma_l}\}_{m,l=1}^k$ with shape parameters $\sigma_l = \min_{r=1, \dots, k, r \neq l} \frac{1}{2} \|\mathbf{K}_r - \mathbf{K}_l\|_2$; more details can be found in [12].

A result of the 3D scene calibration, including the camera and r.b.f. structure calibration is depicted in Figure 6a.

2.5 3D Motion Estimation

Facial expressions are 3-dimensional as they are produced by 3D deformation of the skin, the face shape (due to articulation) as well as the head movements

(3D rigid motion of the head). Extracting these 3D deformation parameters is an ill-posed problem. Indeed, due to the perspective projection only the apparent 2D motion could be recovered, as the third dimension of the motion and structure has been lost after projection. Without the knowledge of the face surface from range or stereo information [13], or without a coarse face structure and an accompanying motion model these 3D motions cannot be estimated.

Describing 3D face motion has its roots in computer graphics. Parke et al. [14] pioneered the modeling of face gestures by parameterizing a 3D human face model. This is done by introducing parameters that each move a group of vertices. The knowledge about the human face movements is thus contained in the specification of the face model. This direct parametrization of the geometry of the face model is used as a constraint in extracting face motion from an image sequence [15]. The estimation method considers the amount of adherence to the brightness constraint equation (related to optical flow) as a discrepancy measure between the observed image irradiance changes and the projected face model changes. That measure is minimal for the true motion model's parameters. This automatic tracking of the face and facial features yields information that can also be used for describing facial expressions by MPEG4 video compression standardization's Facial Animation Parameters (FAP). Within the video coding framework, many researchers followed the optical-flow-based FAP extraction approach [16] [17].

Instead of estimating directly the face geometry, the underlying physical mechanisms of the deformations can be extracted. A mass-spring model is built by replacing the connections between the vertices of the wire-frame by springs and by appointing a point mass to each vertex. Waters [18] modeled face muscles as geometric deformations of a face model with a nonlinear geometric interpolation method to approximate the mass-spring behavior. In facial animation systems that employ morphing of 3D models, the layered animation approach is still being improved [3]. Lee, Terzopoulos, and Waters [19] were the first researchers to apply the dynamic mass-spring system to facial modelling. The amount of 2D movements (with respect to a neutral face) of highlighted eyebrows and mouth and nose furrows of a person estimated in the image sequence was used as a weight of the 3D muscle spring contraction. Performance-driven animation was achieved with a two-layer mass-spring skin model. Following such physics-based approach, Essa [20] estimated a set of muscle parameters from the image motion field. The optical flow measurements in the image sequence were coupled to the deformations in depth of a finite element face model by using the convex nature of the face (via a spherical wrapping). The estimation of the muscle values was incorporated in a control framework of a dynamic system, using a continuous time Kalman filter.

In our research, we propose to employ a mechanically-based finite element model of face deformation inside the optical-flow-based motion estimation formulation. The face structure is given as the calibrated 3D face wire-frame, and a

rigid and nonrigid natural motion model is employed to estimate the 3D motion of the face from a single image sequence measurement. The motion of the 3D face model that best accounts for the entire observed flow is interpreted as the head motion and can be used to create the gestures of a 3D virtual face. This is done by relating the 3D apparent velocity field, called scene flow \mathbf{W} [21], to the optical flow \mathbf{u} in successive images via equation (3), for which the brightness constancy constraint (4) is valid [22]:

$$\mathbf{u} = (u, v) = \mathbf{J}_{\mathbf{X}} \mathbf{W} \quad \text{with } \mathbf{J}_{\mathbf{X}} \triangleq \frac{\partial \mathbf{x}}{\partial \mathbf{X}} = \frac{f}{z} \begin{pmatrix} 1 & 0 & -\frac{x}{z} \\ 0 & -1 & \frac{y}{z} \end{pmatrix} \quad (3)$$

$$I_i u + I_j v + I_t = 0 \quad (4)$$

where $\mathbf{J}_{\mathbf{X}}$ is the scene flow projection Jacobian; $\mathbf{X} = (x, y, z)$ are the 3D coordinates; $\mathbf{x} = (i, j)$ are the image coordinates of a point ; and I_i, I_j, I_t are the gradients of the image I in i - and respectively j - and t -direction.

The face motion will be determined as the 3D face model's motion and estimated by registering the projection of the model's parameterized natural scene flow \mathbf{W} to the measured optical flow \mathbf{u} . For details about the implementation of the approach, we refer the reader to the PhD-thesis [12]. We will thus select the parameters for which this projected scene flow, denoted as *modeled optical flow* $\tilde{\mathbf{u}}(\mathbf{W})$, can most likely resemble the measured optical flow in the least-squares sense, written as

$$\widehat{\mathbf{W}} = \arg \min_{\mathbf{W}} \left[\left(\sum_{k=1}^m \|\mathbf{u}_k - \tilde{\mathbf{u}}(\mathbf{W}_k)\|^2 \right) + \psi \right] \quad (5)$$

This parameterized flow \mathbf{W} consists of a bulk rigid motion \mathbf{V} , including all apparent 3D rotations and translations of the face model, and of the nonrigid displacements \mathbf{U} of the soft skin face tissue caused by muscle forces. Regularization terms ψ , required to solve the ill-posed 3D recovery problem, can now be applied on the scene using the physical constraint of smoothly varying muscles forces that lie tangential to the face surface.

Our approach gives several improvements to the previously mentioned state-of-the-art methods, namely:

- The optical flow in the image to attain with the motion model is back-projected to 3D face model motion considering the perspective projection.
- As compared to the mass-spring models in the state-of-the art [19], forces on the face are seen as distributed muscle loads and thus no face model muscle topology has to be determined. This is realized by implementing the modeled optical flow using a finite element face model that is solved for the thin shell skin surface displacement when a distributed muscle load is applied. Furthermore, the face motions are not restricted to geometrical transformations as in [15,16,17,23].
- Regularization of the solution is foreseen by physically-based restrictions of the distributed muscle forces.

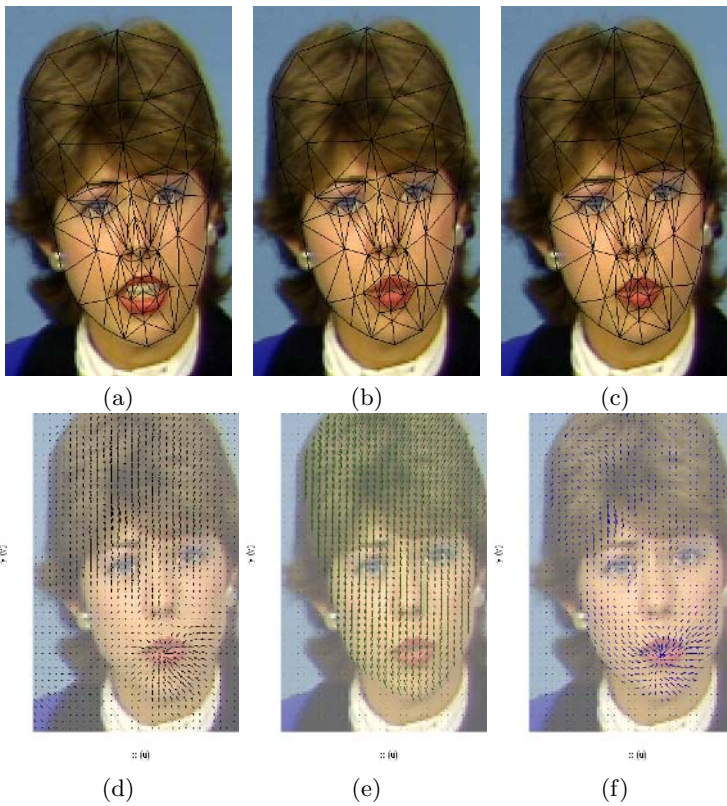


Fig. 6. (a) calibrated face model; (b) rigid motion and (c) nonrigid motion applied on the face model; (d) measured optical flow (from the image in a to b); (e) estimated rigid 2D flow; (f) remaining nonrigid 2D flow



Fig. 7. *Peter* sequence [16] projected model with applied 3D nonrigid motion from estimation, for 4 consecutive frames

Animating the face model with the extracted 3D displacement, results in a faithful reproduction of the observed gestures. Results of the rigid and nonrigid motion estimation are shown in Figure 6b-f for the *Claire* sequence. The estimation of the mouth opening in the *Peter* sequence is displayed in Figure 7.

2.6 MPEG4 Animation

The illusion of motion of a 3D virtual face is created by changing the shape of its visual surface. The parameterized control of storage-efficient geometrical changes is defined in the MPEG4 video compression standard [1,2]. A Facial Animation Parameter (FAP) encodes the normalized magnitude of displacement of a feature point along one axis direction and is associated with the movement of a key face zone. Such a local face animation technique is more flexible than the 'face space' animation in which a database of normalized face models restricts the allowed deformations [3]. The rules of application of the FAP on a wire-frame face model are specified by the non-normalized *FacDefTables*, encoding the displacements of the vertices in each FAP's key zone. By specifying these tables only for the feature points, and using the 3D calibration for the other face points, all FAPs are dynamically applied together on the face model in a linear way:

$$\mathbf{L}_m = \mathbf{K}_m + \sum_{a=1, \dots, g} \text{FAP}_a \cdot \text{FAPU}_a \cdot D_{ma} \cdot \mathbf{pos}_a \quad m = 1, \dots, k \quad (6)$$

$$\Rightarrow \mathbf{Y}_n(\mathbf{FAP}) = \mathbf{f}(\mathbf{A}, \mathbf{B}, \mathbf{K}, \mathbf{N}_n) + \mathbf{FAP} \mathbf{h}(\mathbf{FAPU}, \mathbf{D}, \mathbf{A}, \mathbf{B}, \mathbf{K}, \mathbf{N}_n) \quad n = 1, \dots, N \quad (7)$$

where g is the number of deformation FAPs working on feature point \mathbf{K}_m to obtain \mathbf{L}_m ; the \mathbf{FAPU} convert the displacement values to the face model's unit system; the displacement table \mathbf{D} has elements D_{ma} describing how much a feature point moves \mathbf{K}_m in an axial direction \mathbf{pos}_a when FAP_a is applied; and $\mathbf{Y}_n(\mathbf{FAP})$ is the new position of the n -th vertex of the model after FAP application on the neutral face vertex \mathbf{N}_n using (6) in the r.b.f. interpolation in (2) to build the functions \mathbf{f} and \mathbf{h} .

This procedure considerably reduces the amount of data to be stored in the rule-tables for animation, and consequently the design face animations becomes more easy. Estimating the FAP values that best comply to the previously extracted 3D nonrigid face motion is cast as a linear least-squares problem (8) which results in a semantical representation of the face gestures.

$$\widehat{\mathbf{FAP}} = \arg \min_{\mathbf{FAP}} \sum_{n=1}^N \|\mathbf{Y}_n(\mathbf{FAP}) - (\mathbf{N}_n + \mathbf{U}(\mathbf{N}_n))\|^2 \quad (8)$$

where $\mathbf{U}(\mathbf{N}_n)$ is an estimated 3D displacement vector of a (nonrigid) expression of the neutral face.

The integrated analysis and the synthesis provided by VRML-like viewers to show MPEG4 scenes flexible enough to create simple to sophisticated animated faces.

3 Conclusion

The FASS has achieved performance-driven animation by estimating several face parameters from a single recorded face image sequence. The integration of the techniques of color and shape segmentation, kernel-based tracking, image filtering, statistical notions, semantical descriptions, 3D to 2D projection, natural motion models, and the MPEG4 representation provide a well-founded extendible framework that is capable to meet the specific needs of gesture communication via faces.

References

1. ISO/IEC: Jtc 1/sc 29/wg 11 n2501 coding of moving pictures and audio, information technology - generic coding of audio-visual objects, part1:systems, final draft international standard. Atlantic City (1998)
2. ISO/IEC: Jtc 1/sc 29/wg 11 n2502 coding of moving pictures and audio, information technology - generic coding of audio-visual objects, part2:visual, final draft international standard. Atlantic City (1998)
3. Ostermann, J., Weissenfeld, A.: Talking faces - technologies and applications. In: 17th International Conference on Pattern Recognition(ICPR'04), Cambridge UK, August 23 - 26, 2004. Volume 3. (2004) 826 – 833
4. Ravyse, I., Enescu, V., Sahli, H.: Kernel-based head tracker for videophony. In: The IEEE International Conference on Image Processing 2005 (ICIP2005), Genoa, Italy, 11-14/09/2005. Volume 3. (2005) 1068–1071
5. Ikeda, O.: Segmentation of faces in video footage using hsv color for face detection and image retrieval. In: The IEEE International Conference on Image Processing (ICIP), Barcelona Spain, 14-17/9/2003. Volume 3. (2003) 913–916
6. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: IEEE Computer Science Conference on Computer Vision and Pattern Recognition 2001, CVPR2001, December 08 - 14, 2001, Kauai, Hawaii. Volume 1. (2001) 511–518
7. Zivkovic, Z., Krse, B.: An em-like algorithm for color-histogram-based object tracking. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04), Washington, D.C., USA, June 27 - July 02, 2004. Volume 1. (2004) 798–803
8. Matthews, I., Bangham, J., Harvey, R., Cox, S.: A comparison of active shape model and scale decomposition based features for visual speech recognition. In: Proceedings European Conference on Computer Vision: Lecture Notes in Computer Science (ECCV), Freiburg. (1998) 514–528
9. Ravyse, I., Sahli, H., Reinders, M., Cornelis, J.: Eye activity detection and recognition using morphological scale-space decomposition. In: 15th International Conference on Pattern Recognition, ICPR2000 (September 3-8, 2000) Barcelona, Spain. Volume 1. (2000) 1080–1083
10. Pighin, F., Hecker, J., Lischinski, D., Szeliski, R., Salesin, D.H.: Synthesizing realistic facial expressions from photographs. In: SIGGRAPH 98, in Computer Graphics Proceedings, Annual Conference Series. (1998) 75–84
11. Schaback, R.: Creating surfaces from scattered data using radial basis functions. In: Mathematical Methods for Curves and Surfaces, in Computer Aided Geometric Design III. M. Daehlen, T. Lyche and L.L. Schumaker (eds.), Vanderbilt University Press, Nashville (1995) 477–496

12. Ravyse, I.: Facial Analysis and Synthesis. PhD thesis, Vrije Universiteit Brussel, Dept. Electronics and Informatics, Belgium (2006) online: www.etro.vub.ac.be/Personal/icravyse/RavysePhDThesis.pdf.
13. Vedula, S., Baker, S., Rander, P., Collins, R., Kanade, T.: Three-dimensional scene flow. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27** (2005) 137–154
14. Parke, F.I., Waters, K.: *Computer Facial Animation*. A K Peters (1996) ISBN 1-56881-014-8.
15. Li, H., Lundmark, A., Forchheimer, R.: Image sequence coding at very low bitrates: A review. *IEEE Transactions on Image Processing* **3** (1994) 589–605
16. Eisert, P.: *Very Low Bit-Rate Video Coding Using 3-D Models*. PhD thesis, Universität Erlangen, Shaker Verlag, Aachen, Germany (2000) ISBN 3-8265-8308-6.
17. Yilmaz, A., Shafique, K., Shah, M.: Estimation of rigid and nonrigid motion using anatomical face model. In: *Int. Conference on Pattern Recognition, Quebec, Canada ICPR2002*. (2002)
18. Waters, K.: A muscle model for animating three-dimensional facial expression. *Computer Graphics ACM* **21** (1987) 17–24
19. Lee, Y., Terzopoulos, D., , Waters, K.: Constructing physicsbased facial models of individuals. In: *Proceedings of the Graphics Interface '93 Conference, Toronto, ON, Canada*. (1993) 1–8
20. Essa, I., Basu, S., Darrell, T., Pentland, A.: Modeling, tracking, and interactive animation of faces and heads using input from video. In: *Computer Animation '96, Geneva, Switzerland*. (1996)
21. Spies, H., Jähne, B., Barron, J.L.: Range flow estimation. *Computer Vision Image Understanding (CVIU2002)* **85** (2002) 209–231
22. Klaus, B., Horn, P.: 12. Motion Field and Optical Flow. *The MIT Electrical Engineering and Computer Science Series*. In: *Robot Vision*. MIT - Press (1986) ISBN 0-262-08159-8.
23. Torresani, L., Yang, D.B., Alexander, E.J., Bregler, C.: Tracking and modeling non-rigid objects with rank constraints. *IEEE CVPR 2001 Best Student Paper Award* (01)

Detection of Pathological Cells in Phase Contrast Cytological Images

Marcin Smereka¹ and Grzegorz Glab²

¹ Institute of Computer Engineering, Control & Robotics,
Wroclaw University of Technology, Janiszewski Str. 11/17,
50-372 Wroclaw, Poland

Marcin.Smereka@pwr.wroc.pl

² Gynecological Clinic GMW, 1st Maj Str. 9/72,
Opole, Poland

Abstract. This paper presents a practical combination of image processing and pattern recognition techniques in order to identify pathological and atypical cells in phase contrast cytological images. The algorithms involved in the processing cover: oriented edge detection, ridge following, contour grouping and ellipse fitting. The Hough Transform and other techniques are discussed for comparison. Various pattern recognition techniques are tested and compared. All the exploited algorithms were customized to reflect specificity of phase contrast images and apriori-knowledge of cytological smear. Possible applications of this algorithm for automated screening systems are enumerated.

1 Introduction

The diagnostic cytology is an integral component of gynecological examinations. It enables early detection of precancerous lesions in the uterine cervix. The phase contrast (Ph) microscopy is relatively a new technology in this area. Traditional Pap-Smear tests require staining and fixing, what usually takes several days before the diagnosis is issued [1]. Ph microscope allows for immediate diagnosis and completes the result of colposcopic examination [2]. Computer aided processing of Ph images can additionally improve performance and quality of examination. The joint research conducted at Wroclaw University of Technology and Gynecological Clinic GMW in Opole is aimed at recognition and classification of objects present in the images. It is our belief that computer image processing system could effectively perform the preliminary screening, and provide only pathological and atypical cells for physician's evaluation. Interesting medical objects occur rarely in early cancer phases, so the system could prevent from monotonous inspecting hundreds of microscopic images by a human.

1.1 Oncologic Classification of Cells

Cervico-vaginal smears give a gynecologist a lot of information about actual hormonal balance and vaginal biocenosis. The major benefit is the possibility of

detecting and controlling the intraepithelial squamous and endocervical neoplasia. A carcinogenesis on uterine cervix is a multi-step and long-lasting process, which can be observed in morphologic cellular changes. After 50 years of evolution of cyto-oncologic knowledge, symptoms of carcinogenesis were divided as follows (Bethesda System — modification 2001) [3]:

- Atypical Squamous Cells of Undefined Significance (ASCUS);
- Low Squamous Intraepithelial Lesion (L-SIL);
- Atypical Squamous Cells — High SIL (ASC-HS);
- High Squamous Intraepithelial Lesion (H-SIL);
- Invasive Plane Carcinoma;
- Atypical Glandular Cells (AGC) and Adenocarcinoma Endocervicale in Situ (AIS);
- Invasive Adenocarcinoma Endocervicale and Endometriale.

Through analysis of the specified features of the squamous cells the smear can be classified as normal, atypical or pathological one.

1.2 Identification of Cell Nuclei

It is assumed that the most significant objects in Ph images are cell nuclei. The size and the shape of a nucleus brings a lot of information about precancerous lesions in progress. Large, irregularly outlined and not uniformly filled nuclei are suspected to be atypical (ASCUS) or pathological ones (L-SIL, ASC-HS, H-SIL)[2]. The goal of the algorithm is to detect cell nuclei and classify them as normal or atypical/pathological. The cell nuclei identification algorithm is based on the following assumptions:

- the Ph microscopy emphasizes edges of objects [4], therefore nuclei detection is equivalent to searching for their boundaries;
- the shape of nucleus is circular or elliptic, so oval patterns are of particular interest;
- the image magnification is known in advance, so the objects within the specific range of radii $[r_{min}, r_{max}]$ are considered only.

2 Related Work

Many automated screening systems have been developed for stained and fixed cytological smears. Techniques are mostly based on color and texture information, therefore they are not applicable to Ph images. Respecting assumptions made, the algorithm should be based on geometric features and detect objects with oval boundaries.

The problem of detecting oval shapes has been extensively studied in literature. The Hough Transform (HT) has been recognized as a very powerful method to detect parametric curves in images [5]. It relies on voting process that maps image edge points into manifolds in an appropriately defined parameter space.

Peaks in the parameter space correspond to detected curves. The direct HT method for detecting ellipses is computationally expensive due to the multidimensional parameter space. Many improvements have been proposed to make these methods more efficient [6,7] and more robust to irregularities [8]. Improved HT was successfully applied to detect regular cell nuclei from Ph cytological images [9], but it hardly dealt with pathological cells.

Active Contours (snakes) are designed to detect objects with boundaries not necessarily defined by gradients [10]. The basic idea is to evolve a curve, subject to constraints from a given image. The initial curve moves governed by an appropriately designed energy function until it stops at the local optimum. The final curve is assumed to form the object boundary. Ray and Acton showed that active contours can also be employed for tracking of moving objects [11]. The energy function calculates the difference between features of object and features of background, so it is useful for stained smears. In the case of Ph images, it is difficult to define features that distinguish nuclei and a cytoplasm clearly.

Another group of methods relies on converting gray-scale image to binary image using edge detection techniques and calculating numerical shape descriptors. Peura and Ilvarinen studied some of Simple Shape Descriptors [12]. The descriptor known as elliptic variance is especially useful for detecting ellipses. Rosin proposed other simple descriptors (moment invariants, Euclidean distances) that can be adapted to measure ellipticity of shapes [13]. Pilu and Fitzgibbon were first who presented a direct method for fitting ellipses to the set of points in the least square sense [14]. Their method is used as a part of the segmentation algorithm presented in this work. Previous methods used a generic conic fitting or an iterative approach to recover elliptic solutions. A variety of 'error of fit' functions have been discussed by Rosin [15].

Low level edge detection operators do not guarantee the generation of continuous boundaries of objects. This makes many image analysis tasks difficult, especially for noisy images. The aim of contour grouping algorithms is to connect edges that are supposed to be parts of the same object. Contour grouping techniques were concentrated mainly on detecting salient curves [16,17]. Improvements are concentrated on favoring closed [18] shapes rather than long and smooth ones. A contour grouping algorithm was successfully used in [19] to isolate irregular shape nuclei. The present work is a continuation of [19].

3 Method

The proposed image processing technique combines the idea of contour grouping with pattern recognition methods in order to detect cell nuclei and provide diagnostic information. It consists of the following steps:

1. Detect edges and orientation.
2. Follow ridges and remove those with high curvature.
3. Perform grouping, extract features for each group and assign groups to classes.

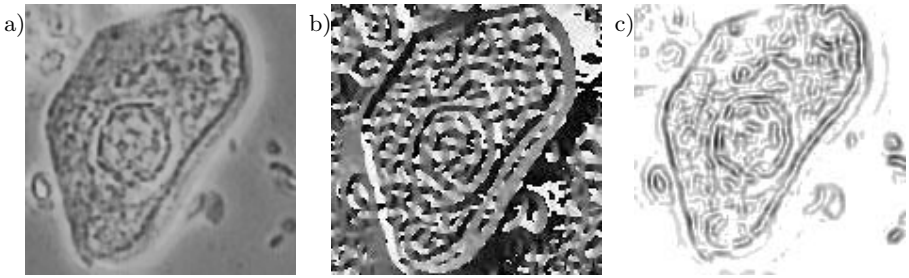


Fig. 1. Example of Ph image processing. a) original image I_I , b) orientation image I_φ encoded by gray levels, c) amplitude image I_A

3.1 Edge Detection

The goal of edge detection is to transform initial gray-scale image I_I (Fig. 1a) to an edge image which assigns amplitude and orientation of edge gradient to each point. To detect an edge orientation and amplitude with a satisfactory precision, I_I image is convoluted with 5×5 horizontal and vertical Prewitt operators G_X and G_Y .

$$G_X^{(x,y)} = y - 3; \quad G_Y^{(x,y)} = x - 3; \quad x, y = 1 \dots 5. \quad (1)$$

As the result horizontal and vertical gradient images are obtained I_X, I_Y .

$$I_X = I_I * G_X; \quad I_Y = I_I * G_Y. \quad (2)$$

To compute orientation and amplitude images I_φ, I_A (Fig. 1b,c), Cartesian coordinates are transformed to polar ones.

$$I_A^{(x,y)} = \sqrt{(I_X^{(x,y)})^2 + (I_Y^{(x,y)})^2}; \quad I_\varphi^{(x,y)} = \text{atan2}(I_Y^{(x,y)}, I_X^{(x,y)}). \quad (3)$$

3.2 Ridge Following Algorithm

A cytoplasm of a cell in Ph image consists of randomly placed spots and highly curved short line segments. These shapes can be easily misclassified as nuclear walls or cell walls. To separate real and phantom edges an algorithm was developed to extract long and smooth edges.

This algorithm finds the initial point (the point with the maximum amplitude), follows the edge in both directions using edge orientation (I_φ) and stops eventually, where the edge amplitude drops below a given threshold t_A . To avoid side effects such as loops and adjacent ridges, it is assumed that the ridge is $w = 5$ pixels wide and the local neighborhood of a pixel is removed from image after the pixel selection (Fig. 2a). The value of w corresponds to the size of edge operators. Pixels that form one edge segment are removed from the image and the process is repeated until no new starting point with amplitude larger than t_A is

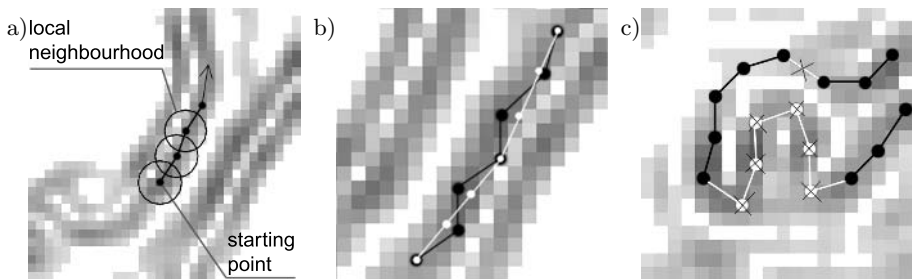


Fig. 2. a) A ridge following, b) 'staircase' and smoothing, c) a segment after removing highly curved points and inflexion points

available. The choice of the optimal value of t_A is a trade-off between sensitivity for weak and soft edges and performance of the grouping process. Experiments shown that the best results were obtained for $t_A = 0.25 \max I_A^{(x,y)}$.

After all segments were created, each segment must be free of the 'staircase' effect. To accomplish this, each quadruple of consecutive points $(i, i+1, i+2, i+3)$ for each edge segment is tested. In the case of inflexion ($infl < 0$):

$$\Delta x_i = x_{i+1} - x_i; \quad \Delta y_i = y_{i+1} - y_i; \quad (4)$$

$$infl = ((\Delta x_{i+1} \cdot \Delta y_i) - (\Delta x_i \cdot \Delta y_{i+1})) \cdot ((\Delta x_{i+2} \cdot \Delta y_{i+1}) - (\Delta x_{i+1} \cdot \Delta y_{i+2})) \quad (5)$$

internal pixels $(i+1, i+2)$ are smoothed linearly (Fig. 2b).

$$\tilde{x}_{i+1} = 1/3 \cdot (x_{i+3} - x_i) + x_i; \quad \tilde{y}_{i+1} = 1/3 \cdot (y_{i+3} - y_i) + y_i; \quad (6)$$

$$\tilde{x}_{i+2} = 2/3 \cdot (x_{i+3} - x_i) + x_i; \quad \tilde{y}_{i+2} = 2/3 \cdot (y_{i+3} - y_i) + y_i. \quad (7)$$

After this operation, coordinates of points have subpixel accuracy. Then the curvature κ_i is calculated for each point.

$$\Delta \tilde{x}_i = \frac{\Delta x_i}{\sqrt{\Delta x_i^2 + \Delta y_i^2}}; \quad \Delta \tilde{y}_i = \frac{\Delta y_i}{\sqrt{\Delta x_i^2 + \Delta y_i^2}}; \quad (8)$$

$$\Delta x_i = x_{i+1} - x_i; \quad \Delta y_i = y_{i+1} - y_i; \quad (9)$$

$$\kappa_i = (\Delta \tilde{x}_i - \Delta \tilde{x}_{i-1})^2 + (\Delta \tilde{y}_i - \Delta \tilde{y}_{i-1})^2. \quad (10)$$

Points with curvature over the threshold value t_κ are removed, therefore segments are split into smaller parts (Fig. 2c). The value $t_\kappa = 2$ was chosen experimentally and corresponds to the angle $\pi/2$. Additionally, inflexion points and very short segments are removed in order to speed up further processing.

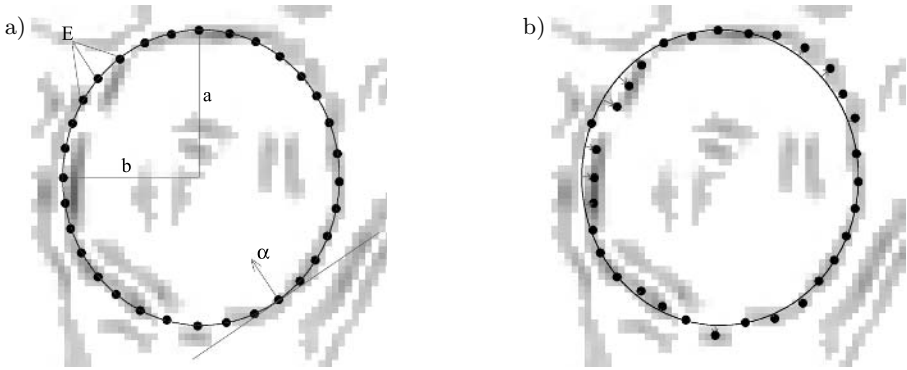


Fig. 3. a) Example of fitting an ellipse to a group of contours; b) The ellipse after displacement

3.3 Grouping

Let us define a graph G , where vertices correspond to edge segments and arcs connect pairs of segments placed close enough to each other. Two segments are close to each other if the distance between their centroids is smaller than $2r_{max}$. A boundary of a nucleus consists of one or more segments. The goal of contour grouping algorithm is to group segments that form the nuclear boundary, so searching for nucleus is equivalent to searching for a path in the graph.

The search algorithm extracts all possible paths in G containing up to 4 vertices. For each tested path the ellipse-fitting algorithm [14] is executed to form the best-fit ellipse to the edge points of the path (Fig. 3a). The boundary of the ellipse is represented by the set E , and its interior – by the set S . Let a and b denote major and minor semi-axis of the ellipse and $\alpha^{(x,y)}$ is a vector normal to the tangent to the ellipse at point (x,y) . Because a nucleus is not ideally elliptic, the points of E are displaced within a limited range along the direction $\alpha^{(x,y)}$. Points are attracted by the nearest pixels of the maximum amplitude (Fig. 3b). This is a heuristic, simplified and low cost version of the active contour technique.

For each ellipse that meets the size constrains:

$$r_{min} \leq a \leq r_{max}; \quad b/a > 0.5 \quad (11)$$

numeric features collected in Table 1 are calculated. Some features were selected to distinguish real object from phantom ones (random set of edges), when the other features were designed to match pathologies. Features are provided to a trained classifier that assigns the ellipse to one of two classes: abnormal nuclei (A_1) or other objects (A_2). Abnormal nuclei are: ASCUS, L-SIL, ASC-HS, H-SIL. The other objects cover normal nuclei, nuclei-like objects and phantom objects. These classes allow to distinguish interesting objects (A_1) from uninteresting ones (A_2). When the grouping process is completed, a set of contour groups classified as A_1 is selected. Each group represents final abnormal nucleus.

Table 1. Extracted features. Symbols \searrow \nearrow denote low/high level, respectively

Name	Formula	Comment
major semi-axis	a	\searrow probably regular, \nearrow probably pathologic
avg. edge amplitude	$\frac{1}{ E } \sum_{(x,y) \in E} I_I^{(x,y)}$	\searrow probably background, \nearrow sharply outlined object, probably nucleus
avg. misorientation	$\frac{1}{ E } \sum_{(x,y) \in E} \alpha^{(x,y)} - I_\varphi^{(x,y)} $	\searrow elliptic object, probably nucleus, \nearrow random edges
coverage	$\sum_{(x,y) \in E} \begin{cases} 1, & I_A^{(x,y)} > t_A \\ 0, & \text{otherwise} \end{cases}$	\searrow random edges, \nearrow elliptic object, probably nucleus
aspect ratio	b/a	\searrow random edges, \nearrow elliptic object, probably nucleus
avg. level of texture	$t = \frac{1}{ S } \sum_{(x,y) \in S} I_I^{(x,y)}$	\searrow probably nucleus, \nearrow probably granulocyte or dust
variance of texture	$\frac{1}{ S } \sum_{(x,y) \in S} (I_I^{(x,y)} - t)^2$	\searrow probably regular, \nearrow probably pathologic

Table 2. The expert classification for experimental images

Description	Quantity Class	
Pathological nuclei	41	A_1
Atypical nuclei	16	A_1
Normal nuclei	389	A_2
Nuclei-like objects	178	A_2
Phantom objects	1812	A_2

4 Experiments

Image processing and feature extraction process was implemented in C++. Classification experiments were performed using WEKA Environment [20]. The experimental set contained 2436 vectors of features classified by a medical expert. The expert classification is given in Table 2. The little portion of A_1 -class objects is caused by its rare occurrence in real images. Training and testing were executed using cross-validation method. The experimental set was divided into 4 folds. In each pass 3 folds were used for training and 1 for testing.

Several known classification methods were examined with its standard WEKA parameters. The following algorithms were used:

NB — Naive Bayes Classifier [21];

LR — Multinomial logistic regression model with a ridge estimator [22];

SLR — Linear logistic regression model [23];

Table 3. Classification performed with different algorithms

Algorithm	p_1 (%)	p_2 (%)	p_t (%)	p_w (%)
VFI	12.28	4.58	4.76	8.43
MP	29.82	0.88	1.56	15.04
NB	29.82	2.98	3.61	16.40
NN	36.84	0.88	1.72	18.86
ADT	40.35	0.55	1.48	20.45
RIP	40.35	1.18	2.09	20.77
kNN(k=5)	43.86	0.33	1.35	22.10
NBT	43.86	0.80	1.81	22.33
LR	49.12	0.34	1.47	24.73
SLR	56.14	0.38	1.68	28.26
J48	56.14	0.71	2.01	28.43
kNN(k=10)	59.65	0.00	1.40	29.83
LMT	59.65	0.46	1.85	30.06
KS	59.65	0.50	1.89	30.08

MP — Multilayer Perceptron trained with backpropagation method (1 hidden layer, 5 hidden nodes, 2 output nodes, Learning Rate=0.3, Momentum=0.2) [24];

NN — Nearest Neighbour Classifier [25];

kNN — k-Nearest Neighbour Classifier [25];

KS — K-Star is an instance-based classifier, that uses an entropy-based distance function [26];

VFI — Classification by voting feature intervals [27];

ADT — Alternating Decision Tree [28];

J48 — A pruned C4.5 decision tree [29];

LMT — Logistic model tree [23];

NBT — Decision tree with naive Bayes classifiers at the leaves [30];

RIP — Repeated Incremental Pruning to Produce Error Reduction (RIPPER) [31];

Number of elements in sets A_1 and A_2 were $n_1 = 57, n_2 = 2379$. m_1, m_2 denotes the number of misclassified objects in A_1 and A_2 , respectively. The result of classification is given in Table 3. Measures for evaluating classifiers were defined as follows: p_1, p_2 are the partial probabilities of misclassification,

$$p_1 = \frac{m_1}{n_1}; \quad p_2 = \frac{m_2}{n_2} \quad (12)$$

while p_t is the total misclassification probability, and p_w is the weighted misclassification probability.

$$p_t = \frac{m_1 + m_2}{n_1 + n_2}; \quad p_w = \frac{p_1 + p_2}{2}. \quad (13)$$

All tested algorithms present very sharp separation for class A_2 , and relatively weak separation for class A_1 . It means, that lots of abnormal nuclei could be

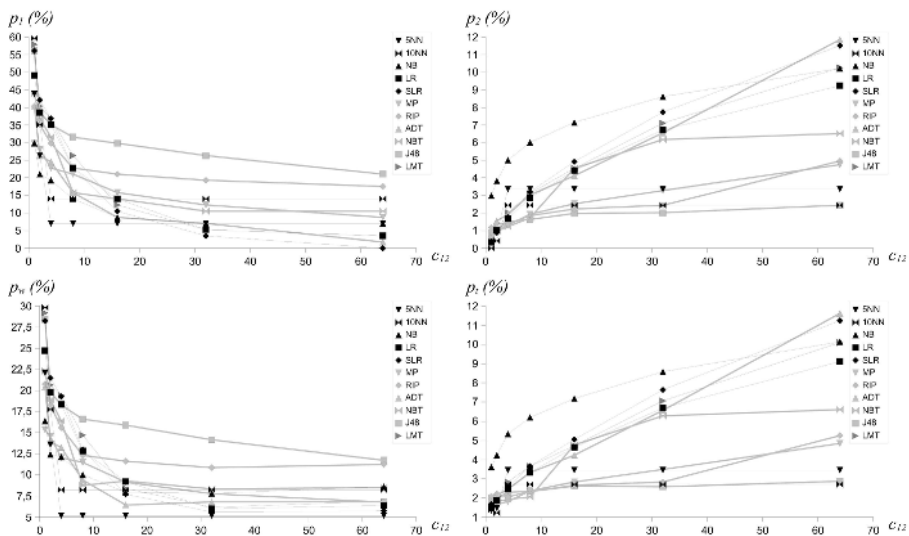


Fig. 4. Measures p_1, p_2, p_w, p_t as the function of c_{12}

Table 4. Cost sensitive classification for cost factor $c_{12} = 32$

Algorithm	$p_1(\%)$	$p_2(\%)$	$p_t(\%)$	$p_w(\%)$
kNN(k=5)	7.02	3.36	3.45	5.19
SLR	3.51	7.73	7.64	5.62
LR	5.26	6.73	6.69	5.99
LMT	5.26	7.10	7.06	6.18
ADT	7.02	6.56	6.57	6.79
MP	12.28	3.28	3.49	7.78
NB	7.02	8.62	8.58	7.82
kNN(k=10)	14.04	2.44	2.71	8.24
NBT	10.53	6.18	6.28	8.35
RIP	19.30	2.44	2.83	10.87
J48	26.32	2.02	2.59	14.17

missed. It is caused obviously by unbalanced training set. It is desirable to achieve better separation for class A_1 , even if the separation for class A_2 deteriorates. This can be achieved by using cost sensitive classifiers or by reweighting the training data. Let C denotes the cost matrix, and c_{12}, c_{21} are the misclassification costs for A_1, A_2 , respectively.

$$C = \begin{bmatrix} c_{11} & c_{12} \\ c_{21} & c_{22} \end{bmatrix}. \tag{14}$$

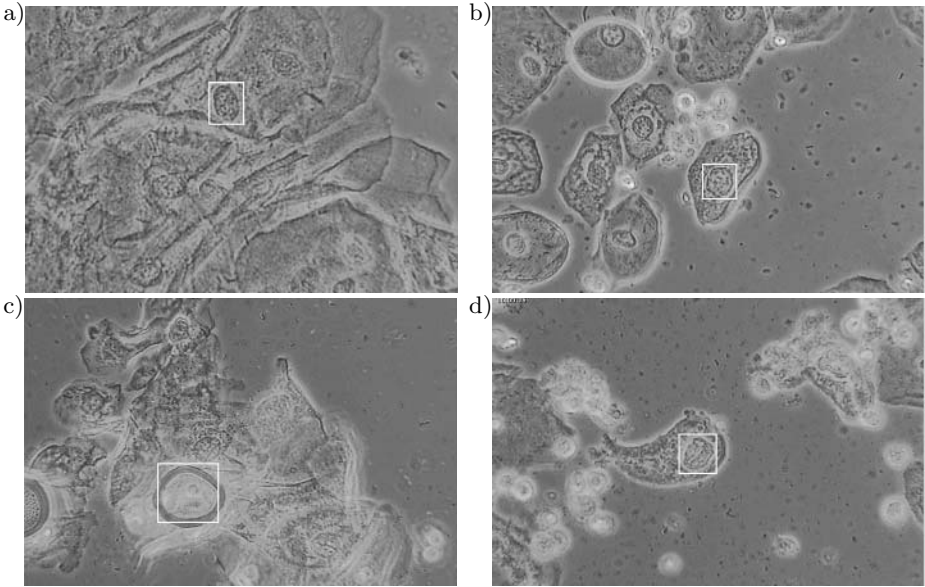


Fig. 5. Exemplary Ph images after processing and visualization

Note, that only a few algorithms allow to minimize weighted cost, so to examine kNN, RIP and J48 instead of minimizing weighted cost, the training data instances are reweighted.

In the next experiments we assume $c_{11} = c_{22} = 0$; $c_{21} = 1$. The relationship between measures p_1, p_2, p_t, p_w and the cost factor c_{12} is illustrated in Fig. 4.

It can be noticed, that due to the increase of the cost factor the separation for class A_1 increased while the separation for class A_2 dropped. To minimize the weighted misclassification probability p_w , we select a constant cost factor $c_{12} = 32$. For most classifiers this value of the cost factor minimized p_w . As one can expect $c_{12} \approx \frac{n_2}{n_1}$. Results of cost sensitive classification for $c_{12} = 32$ are collected in Table 4.

5 Conclusions

In this paper the algorithm was proposed to detect pathological cells in Ph cytological images. The correctness of classification ($1 - p_t$) is ranged between 90% and 95% depending on the applied classifier. The lowest total misclassification p_t is achieved for kNN methods, decision trees and rules (J48, RIP). To preserve low misclassification for class A_1 , the measure p_w was chosen to evaluate cost sensitive classifiers. The kNN algorithm and linear regression algorithms are the leading ones due to their correctness and simplicity. Other methods (LMT, MP, NB) also keep the high correctness of classification. It means that the maximum possible class

separation (90%–95%) is achieved for selected features. Further improvements of the algorithm should concentrate on searching for more distinctive features.

The presented algorithm could be useful for preliminary oncological screening of cytological images because it is relatively fast and robust. Presently, the algorithm is being tested in Gynecological Clinic GMW in Opole, Poland. An image processing stage of one 640x480 image takes from 4 up to 6 seconds on AMD ATHLON 1.8GHz, 512 MB RAM. A classification stage is immediate as the model and its parameters were fixed off-line. Examples of classification using trained SLR classifier are given in Fig. 5. Numerous edge segments resulting from various cell structures did not influence the correctness of detection significantly. Oncologic screening system would rely on taking series of images of microscopic cytological smear. If any abnormal cell is detected in an image, it is exposed for visual inspection by physician. Otherwise, the image is rejected.

References

1. Koss, L.G.: The papanicolaou test for cervical cancer detection: A triumph and a tragedy. *J. Amer. Med. Assoc.* (1996) 737–743
2. Glab, G., Florczak, K., Jaronski, J., Licznarski, T.: *Gynecological cyto-diagnosis in phase contrast microscopy* (in Polish). Blackhorse Publ., Warszawa (2001)
3. Wright, T.C., Kurman, G.J., Ferenczy, A.: *Cervical intraepithelial neoplasia. Pathology of the Female Genital Tract* (1994)
4. Ross, K.F.A.: *Phase contrast and interference microscopy for cell biologists*. Edward Arnold Publ., London (1967)
5. Duda, R.O., Hart, P.E.: Use of the Hough Transform to detect lines and curves in pictures. *Communications of the ACM* **15** (1972) 11–15
6. Atiquzzaman, M.: Coarse-to-fine search technique to detect circles in images. *Int. Journal of Advanced Manufacture Technologies* **15** (1999) 96–102
7. Guil, N., Zapata, E.L.: Low order circle and ellipse hough transform. *J. Pattern Recognition* **30** (1997) 1729–1744
8. Atherton, T.J., Kerbyson, D.J.: Size invariant circle detection. *Image and Vision computing* **17** (1999) 795–803
9. Smereka, M.: Nuclei recognition in phase contrast microscopy images. *Proc. of the 3rd Int. Conf. on Computer Recognition Systems KOSYR'03* (2003) 35–40
10. Chan, T.F., Vese, L.A.: Active contours without edges. *IEEE Trans. Image Processing* **10** (2001) 266–277
11. Ray, N., Acton, S.T., Ley, K.F.: Tracking leukocytes in vivo with shape and size constrained active contours. *IEEE Trans. Med. Imag. (Special Issue on Image Analysis in Drug Discovery and Clinical Trials)* **21** (2002) 1222–1235
12. Peura, M., Iivarinen, J.: Image segmentation using a dynamic thresholding pyramid. *Aspects of Visual Form* (1997) 443–451
13. Rosin, P.L.: Measuring shape: Ellipticity, rectangularity, and triangularity. In *proc. ICPR 2000* (2000) 1952–1955
14. Fitzgibbon, A., Pilu, M., Fisher, R.B.: Direct least square fitting of ellipses. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **21** (1999) 477–480
15. Rosin, P.: Assessing error of fit functions for ellipses. *Graphical models and image processing: GMIP* **58** (1996) 494–502

16. Shashua, A., Ullman, S.: Grouping contours by iterated pairing network. *Neural Info* **3** (1991) 335–341
17. Zhu, Q., Payne, M., Riordan, V.: Edge linking by a directional potential functions (dpf). *Image and Vision Computing* **14** (1996) 59–70
18. Elder, J.H., Zucker, S.W.: Computing contour closure. *ECCV* **1** (1996) 399–412
19. Smereka, M.: Detection of elliptical shapes using contour grouping. *Proc. of the 4th Int. Conf. on Computer Recognition Systems CORES'05* (2005) 443–450
20. Witten, I.H., Frank, E.: *Data Mining: Practical machine learning tools and techniques*. 2nd Edition, Morgan Kaufmann, San Francisco (2005)
21. John, G.H., Langley, P.: Estimating continuous distributions in Bayesian classifiers. *Proc. of the 11th Conf. on Uncertainty in Artificial Intelligence* (1995) 338–345
22. Cessie, S., van Houwelingen, J.C.: Ridge estimators in logistic regression. *Applied Statistics* **41** (1992) 191–201
23. Landwehr, N., Hall, M., Frank, E.: Logistic model trees. *ECML* (2003) 241–252
24. Hertz, J., Krogh, A., Palmer, R.G.: *Introduction to the theory of neural computation*. Addison–Wesley Publ. (1991)
25. Aha, D., Kibler, D.: Instance–based learning algorithms. *Machine Learning* **6** (1991) 37–66
26. Cleary, J.G., Trigg, L.E.: K: An instance–based learner using an entropic distance measure. *Proc. of the 12th Int. Conf. on Machine learning* (1995) 108–114
27. Demiroz, G., Guvenir, A.: Classification by voting feature intervals. *ECML* (1997)
28. Freund, Y., Mason, L.: The alternating decision tree learning algorithm. *Proc. of the 16th Int. Conf. on Machine Learning* (1999) 124–133
29. Quinlan, R.: *C4.5: Programs for Machine Learning*. Morgan Kaufmann Publishers, San Mateo, CA. (1993)
30. Kohavi, R.: Scaling up the accuracy of naive–Bayes classifiers: a decision tree hybrid. *Proc. of the 2nd Int. Conf. on Knowledge Discovery and Data Mining* (1996) 202–207
31. Cohen, W.W.: Fast effective rule induction. *Proc. of the 12th Int. Conf. on Machine Learning* (1995) 115–123

Water Flow Based Complex Feature Extraction

Xin U Liu and Mark S Nixon

ISIS group, School of ECS, University of Southampton, Southampton, SO17 1BJ, U.K.
{x1104r, M.S.Nixon}@ecs.soton.ac.uk

Abstract. A new general framework for shape extraction is presented, based on the paradigm of water flow. The mechanism embodies the fluidity of water and hence can detect complex shapes. A new snake-like force functional combining edge-based and region-based forces produces capability for both range and accuracy. Properties analogous to surface tension and adhesion are also applied so that the smoothness of the evolving contour and the ability to flow into narrow branches can be controlled. The method has been assessed on synthetic and natural images, and shows encouraging detection performance and ability to handle noise, consistent with properties included in its formulation.

1 Introduction

Complex shape extraction is of great interest in practical uses such as vessel detection in iridology. There are two popular techniques which both involve contour evolution: active contours or snakes, and region growing.

Snakes evolve a parameterized curve from an initial position to the boundaries of the object following some rules to minimize a specified energy functional. The functional is defined so that the minimization can give rise to a smooth and even contour. In complex feature extraction, however, the classical snake is of limited use as it needs good initialization near the boundary and cannot handle topological changes like boundary concavities. Many methods have been proposed to overcome these problems. The balloon models [1], distance potentials [2], and gradient vector flow (GVR) field [3] have been introduced as the solutions of initialization and concave boundary detection. Snake energy functionals using region statistics or likelihood information have also been proposed [4, 5]. A common premise is to increase the capture range of the external forces to guide the curve towards the boundaries. For more complex topology detection, several authors have proposed adaptive methods like the T-snake [6] based on repeated sampling of the evolving contour on an affine grid. Geometric active contours [7,8] have also been developed where the planar curve is represented as a level set of an appropriate 2-D surface. They work on a fixed grid and can automatically handle topological changes. However, many methods solve only one problem whilst introducing new difficulties. The balloon models introduce an inflation force so that it can “pull” or “push” the curve to the target boundary, but the force cannot be too strong otherwise “weak” edges would be overwhelmed. Region-based energy can give a large basin of attraction and can converge even when the explicit edges do not exist but it cannot yield as good localization of the contour near the boundaries as edge-based

methods. Level set methods detect complex shapes well at the cost of increased dimensionality and hence much greater complexity.

The region growing techniques mainly rely on the assumption that adjacent pixels in the same object or region have similar characteristics such as intensity and texture. They test the statistics inside the growing region and then decide whether or not the adjacent pixel can be merged according to the specified homogeneity criterion. Region growing techniques are free of topological changes since they are pixel-wise techniques without smoothness constraints [9]. However, this property also tends to yield irregular boundaries and small holes, especially for noisy images [10]. Besides, the region statistics comparison standards on which they are based can lead to inaccurate contour detection.

This paper proposes a new feature extraction method based on water flow. Unlike the famous watershed method, which is based on mathematical morphology and is often combined with snakes [11] and region growing [12], the focus is now on the “water” itself rather than the “landscape” of images because the properties of water, like fluidity and surface tension, are well suited to complex feature extraction. We first introduce the related physical principles and the framework of the technique, and then define all the analogical factors. Finally, results both for synthetic and for real iris images are presented, which show the resolution of problems like topological changes, and good noise immunity.

2 Methodology

Water flow is a compromise between several factors: the position of the leading front of a water flow depends on pressure, surface tension, adhesion/capillarity. There are some other natural properties like turbulence and viscosity, which are ignored here. Image edges and some other characteristics that can be used to distinguish objects are treated as the “walls” terminating the flow. The final static shape of the water should describe the related object’s contour.

The flow is determined by pressure and the resistance. The relationship between the flow rate f_r , the flow resistance R and the pressure difference, is given by:

$$f_r = P_i - P_o \quad (1)$$

where P_i and P_o are pressure of the inflow and outflow, respectively. The pressure difference drives the flow and

$$f_r = AV_{effective} \quad (2)$$

where A is the cross-sectional area and $V_{effective}$ is the effective flow velocity. Hence the velocity can be related to force and resistance through equations (1) and (2).

There are small discontinuities or weak regions existent on the contours which may lead to “leakage” of water. The surface tension, which can form a water “film” to bridge gaps, is then applied to overcome the problem. An attractive force existing between water and walls, named adhesion is defined as the attractive force generated by image edges. It is adopted in the new technique to assist surface tension to bridge edge gaps and allow flow into narrow braches.

2.1 Framework of the Operator

The method has little dependence on the starting contour shape. The only limitation on initialization is that it cannot cross the target object's boundaries. One pixel in the image is considered to be one basic unit of the water, and the pressure between an element and each of its neighbors is assumed to be the same. An adaptive source is assumed so that the water can keep flowing until stasis, where flow ceases. An inner element with symmetrical distribution of neighbors hence suffers zero resultant pressure. A water contour element, however, has asymmetrically distributed neighbors (and possibly an additional adhesive force), thus has non-zero pressure difference which leads to a non-zero velocity by equations (1) and (2), and is possible to move outwards. Hence only boundary elements are of interest.

The flow process is assumed to be made up of two separable steps. The first stage is *acceleration*: the contour element achieves a velocity due to the presence of the pressure difference (and any adhesive force), and the ultimate value is given by equation (1) and (2). The next step is *external movement* where the moving element is now free from the influence from other water elements and suffers only external image forces. This is not consistent with a real action but is sufficient for the digital image analogy and greatly simplifies the algorithm.

The water element can move outwards in any direction for which the component of velocity is positive. However, only if the velocity in the direction is sufficiently large, can the element break through the image resistant forces and reach the new position. To reconcile the flow velocity with forces, dynamical formulae are used. We may compute the displacement of a contour element on each possible direction within a fixed time interval, which is similar to snake techniques. However, for simplicity and avoiding the interpolation problem, a framework like region growing and the greedy snake is used: the element will flow to some positions if certain conditions or formulae are satisfied. Here, an equation describing the *conservation of energy* is employed. If assuming that an element, which has a positive velocity v on a particular direction and is acted by the force F during the process, can arrive at the direction-related position ultimately, then this equalization must be fulfilled:

$$mv_F^2/2 = FS + mv^2/2 \quad (3)$$

where v_F is the final scalar velocity after fixed displacement S and m is the assumed mass. In this equation, force F is a scalar which is positive when the force is consistent with velocity v , and negative otherwise. The summation on the right hand side is just the movement decision operator: only if F is negative, can the summation be negative and thus the equality above cannot be satisfied.

2.2 Flow Driving Force with Surface Tension

The pressure on each contour element should be outwards normal to the contour line. Since the flow on each possible direction will be examined separately, the related component of forces rather than the composition is of interest, and a simple convolution method is used. The force on a contour element is determined by the surface interior i.e., the amount and position of the adjacent elements. For a certain direction, the-re will be supporting and opposing elements. The property is determined by their relative



Fig. 1. The convolution masks for component forces on the direction of (from left to right) 90 degrees, 0 degree, 135 degrees, and 45 degrees. The other 4 masks are the transpose of these and we can define the driving force on the opposite direction as negative.

position to the flow direction. The elements located at the normal to the direction do not affect the movement. The ones located at the inner half exhibit positive effects on the flow and the opposite ones give negative forces since the interactive force between elements is repulsion. The 3x3 templates are shown by figure 1. A matrix **W** is used to save the water information where a water element has value one and others are all zeros. Denoting the convolution template for direction *i* as **T_i**, the corresponding matrix saving the normalized driving force strength on direction *i*, **F_{D,i}**, is then calculated by convolution as:

$$F_{D,i} = W * T_i / S_{PM} \tag{4}$$

where *S_{PM}* is the possible maximum of the convolution sums. For each mask, the maximal value is achieved when water elements locate at all positions of 1's and none of those of -1's and hence *S_{PM}* = 3 for the above masks. The driving force strength on direction *i* at point (*x*, *y*) is then just the (*x*, *y*)th entry of **F_{D,i}**.

The convolution mechanism allows situations of more than two adjacent contour elements which is common in complex shape extraction. The mask size can be expanded so that more information of local water structure can be involved and the calculation will be expected to give a more reasonable result. For instance, a single line would have smaller driving force when using 5*5 mask than that by a 3*3 one because the possible maximum is much larger but the convolution sum is just increased by 1. In addition, the method makes the application of surface tension more straightforward. From physics, the surface tension is decided by the temperature and the water itself. In this image analogy, it is defined as a constant *attractive* force between the contour elements. So in the previous convolution, we can just modify the water matrix with the contour position information so that the point will exhibit attractive forces. This is done by setting contour elements entries in **W** as fixed negative values, like -t. Then, replacing **W** in equation (5) with the new matrix **W'** and noting that the possible maximum is now (3+2t) will give the driving force combined with surface tension. Here, we set t=1.

2.3 Resistance to Flow and the Velocity

From equation (1) and (2), the flow velocity is inversely proportional to the resistance of water. In a physical model, the *flow resistance* is decided by the water, the flow channel and temperature etc. Since this is a physical analogy which offers great freedom in selection of parameter definitions, we can assign high resistance values for unwanted image attributes and low values to preferred ones. For instance, in vessel detection in images of the retina, if the vessels have relatively low intensity, we can define the resistance to be proportional to the intensity of the pixel. If we couple the

resistance with the edge information, the process will become adaptive. That is, when the edge response is strong, resistance would be large and so the flow velocity would be weakened. According to equation (3), the movement decision will now be dominated by the force acting during the exterior movement. Thereby, even if the driving force set by users is too “strong”, the resistance would lower its influence at edge positions and the problem in balloon models [1], where strong driving forces may overwhelm “weak” edges, can be eliminated. From equations (1) and (2), the velocity is:

$$\mathbf{V}_i = \mathbf{F}_i / AR \quad (5)$$

where \mathbf{V}_i is the resulting flow velocity. The direction of \mathbf{V} is the same as the force \mathbf{F}_i . In this paper, A is set as a constant, and R at position (u, v) is determined by

$$R(u, v) = \exp\{-k \mathbf{E}(u, v)\} \quad (6)$$

where \mathbf{E} is the edge response matrix and k controls the fall of the exponential curve.

2.4 Image Forces

The gradient of an edge response map is often defined as the potential force in active contour methods since it gives rise to vectors pointing to the edge lines [3]. This is also used here. The force is large only in the immediate vicinity of edges and always pointing towards them. The second property means that the forces at two sides of an edge have opposite directions. Thus it will attract water elements onto edges and prevent overflow. The potential force on a contour element (x_c, y_c) is given by:

$$\mathbf{F}_p = \nabla \mathbf{E}(x_i, y_i) \quad (7)$$

where $\nabla \mathbf{E}$ is the gradient of the edge map, and (x_i, y_i) are the coordinates of the flow target because the potential force is presumed to act during the second stage of flow where the element has left the contour and is moving to the target.

Adhesion is defined as the attraction between water and adjacent vessel walls in physics. In the image analogy, it is determined by potential force based on an edge map with “flooded” positions set to zero. In this map, the edges that have been occupied by the water are ignored so that the edges are clipped. As water flows, vectors (forces) pointing from the flooded edges to the existent ones are generated iteratively and thus assist in flow to the reserved edge lines. It is defined as

$$\mathbf{F}_A = \nabla \mathbf{D}(x, y) \quad (8)$$

where \mathbf{D} is the edge map eroded by the flowing water and (x, y) are the coordinates of a contour point. This equation effectively defines the attractive force from edges to the water. Therefore, even if the water has flowed onto an edge point, it can still move to the adjacent edges. This will thus help water flow into narrow branches, and “flood” small noise pixel clusters to give noise robustness.

The forces defined above work well as long as the gradient of edges pointing to the boundary is correct and meaningful. However, as with corners, the gradient can sometimes provide useless or even incorrect information. Unlike the method used in the inflation force [1] and T-snake [6], where the evolution is turned off when the intensity is bigger than some threshold, we propose a *pixel-wise* regional statistics based

image force. The statistics of the region inside and outside the contour are considered respectively and thus yield a new image force:

$$F_S = -(\mathbf{I}(x_t, y_t) - \mu_{int})^2 n_{int} / (n_{int} + 1) + (\mathbf{I}(x_t, y_t) - \mu_{ext})^2 n_{ext} / (n_{ext} + 1) \tag{9}$$

where subscripts “*int*” and “*ext*” denote inner and outer parts of the water, respectively; μ and n are the mean intensity and number of pixels of each area, separately; \mathbf{I} is the original image. The equation is deduced from the Mumford-Shah functional [5]:

$$F_1(C) + F_2(C) = \int_{inside(C)} |\mathbf{I}(x_t, y_t) - \mu_{int}|^2 + \int_{outside(C)} |\mathbf{I}(x_t, y_t) - \mu_{ext}|^2 \tag{10}$$

where C is the closed evolving curve. If we assume C_0 is the real boundary of the object in the image, then when C fits C_0 , the term will achieve the minimum. Instead of globally minimizing the term as in [5], we obtain equation (9) by looking at the change of the total sum given by *single* movement of the water element. If an image pixel is flooded by water, the statistics of the two areas (water and non-water) will change and are given by equation (9). The derivation is shown in the Appendix.

The edge-based potential forces can provide a good localization of the contour near the real boundaries (i.e., accuracy) but have very limited capture range and are not suitable for edge corners, whilst the region-based forces have a large basin of attraction but cannot provide good detection accuracy. The complementary properties motivate a unification of the two forces. A convex combination method is hence chosen and the combined force is given by:

$$F = \alpha F_P + (1 - \alpha) F_S \tag{11}$$

where all terms are scalar quantities, and α ($0 \leq \alpha \leq 1$) is determined by the user to control the balance between them.

2.5 Movement Decision Process

Equation (3) has provided the inequality to determine the feasibility of outwards flow. For each contour element, we have presented equations computing driving force \mathbf{F}_D modified by surface tension, adhesion \mathbf{F}_A and resistance R . Flow velocity \mathbf{V} can then be obtained through equation (5). If the velocity points towards the exterior of the water, the element is assumed to leave the original position. A unified image force \mathbf{F} provided by equations (7) and (11) is then turned on. The summation in equation (3) can be computed and the sign determines the result of the movement.

Defining m and S in equation (3) as constants, we can then present the new and detailed expression with parameters defined before:

$$J = \lambda \{ (F_D + F_A) / R(x, y) \}^2 + F \tag{12}$$

where λ is a regularization parameter set by users which controls the tradeoff between the two energy terms. It can be considered to be determined by the combination of mass m , displacement S and area A . Its value reflects smoothing of image noise. For example, more noise requires larger λ . F_A and F_D are the scalar components on the movement direction of \mathbf{F}_A and \mathbf{F}_D , respectively. A positive direction is defined from the origin to the target. The movement decision can be completely made by this operator since the term of right hand side inside the brackets gives the velocity information and J corresponds to the ultimate kinetic energy. If the velocity component is greater than zero and

if J is positive, the movement is said to be feasible and the target point will be flooded by water.

3 Experimental Results

The new technique is applied to both synthetic and natural images, and is evaluated both qualitatively and quantitatively.

3.1 Synthetic Images

There are two sorts of initialization for the method. The water “source” can be either inside or outside the target object. The former is suited to most cases, whilst the latter is useful when simultaneously detecting multiple objects in an image. The whole “background” will be flooded, thus the static water will give the targets’ shape information. An example is given in figure 2 with initialization from the image border. All the shapes are detected including the helical pipe which has boundary concavities. The result is accurate even with some noise contamination.

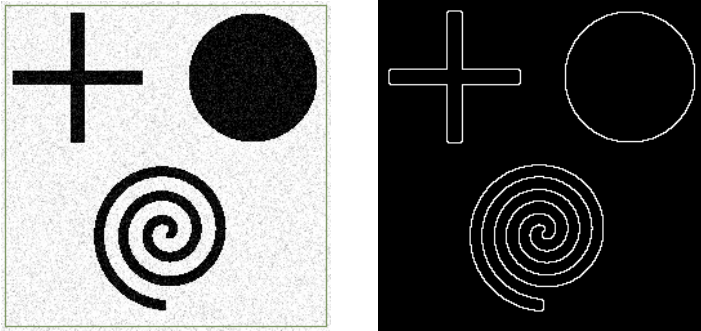


Fig. 2. Multi-object detection: the image is corrupted with 10% Gaussian noise

A 512×512 test image was generated for a performance evaluation according to certain criterion: a) horizontal, vertical and diagonal branches are included; b) narrow and wide branches are presented, respectively; c) there is a circular pipe so that we get a curve with smoothly changing curvature; d) each half of the object has a different intensity so that weak edges exist between them. The image is suited to assess the operator’s ability in complex feature detection, and the noise immunity is also tested by adding Gaussian and impulsive noise to the image. Figure 3 shows the evolutions and the final results. The detection is successful in total, and the immunity to impulsive noise should be emphasized. It’s very difficult for snakes and region growing methods to deal with impulsive noise as the edge response is very strong. In figure 3(c) and (d), almost all the noise points inside the object are flooded, and the detected contour is reasonably accurate. The robustness to impulsive noise arises from the fluidity and the adhesion: water surrounds the small clusters of impulsive noise pixels, and the adhesive force given by the noise response attracts the water to flow over the noise area. So, unless the noise clusters are too large, the noise pixels will be flooded.

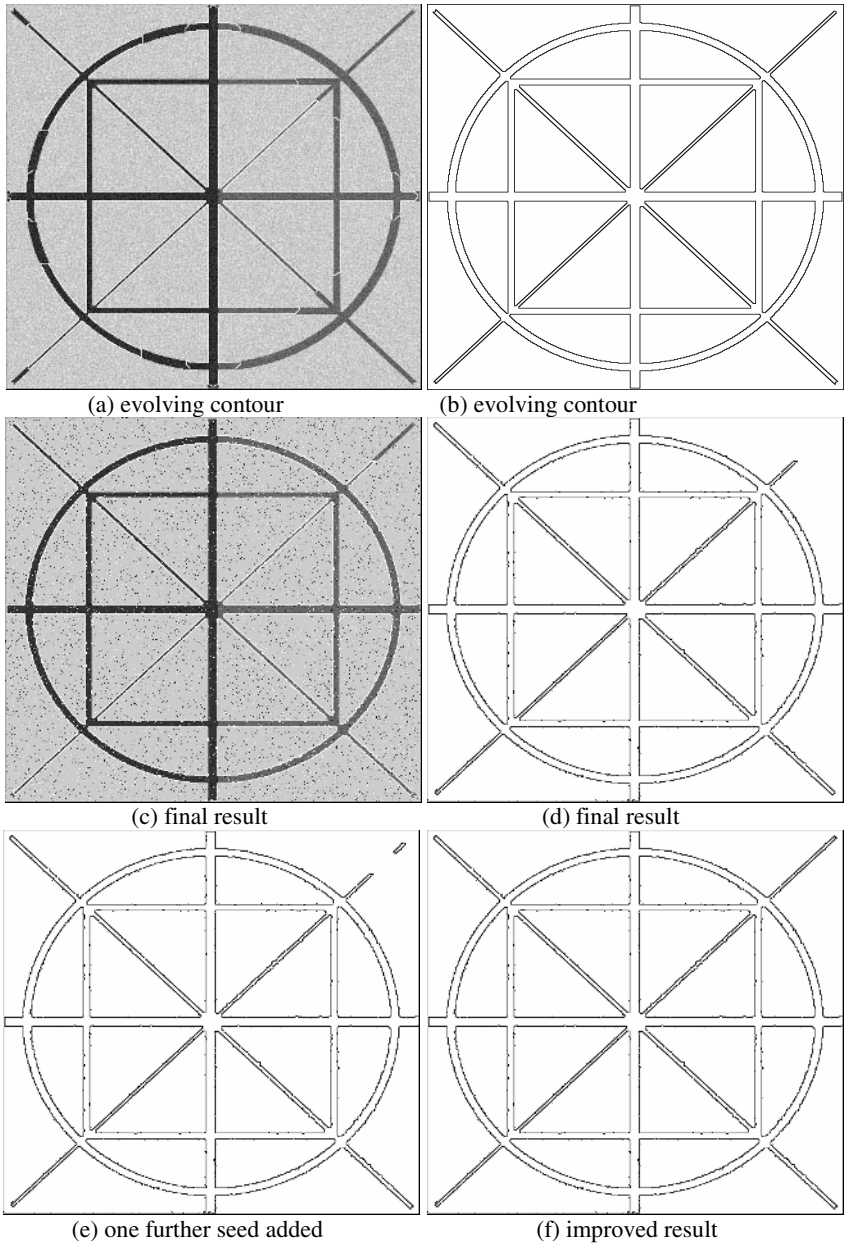


Fig. 3. Flow progress and results for test images contaminated by 10% Gaussian noise (a and b) and 5% impulsive noise (c, d, e and f)

The most significant failure shown by figure 3(d) is the incomplete detection of the thinnest diagonal branch in the top right corner. This is because the branch is too narrow and a noise cluster “blocks” the pipe. In practical applications, this kind of gap will

exist and makes the contour extraction terminate early. Flow from multiple sources can be considered, to overcome the problem. In this simple case, as shown in figures 3(e) and (f), a new source is initialized inside the undetected area. It then fills the region and merges with the “main” part. Therefore a complete detection can be achieved. Similar multiple seeds methods are often incorporated with watershed and region growing techniques, and are not invoked here.

The immunity to noise is also assessed quantitatively, and figure 4 shows the result. The mean square error is used as the criterion with a synthetic test image as the ground truth, which has been deliberately designed to incorporate a narrow boundary concavity. Two typical sorts of noise, Gaussian and impulsive, are added on the image, and the operator performs well and stably for both types of noise, until severe noise contamination (below 10 dB). An example of the detection result is also shown.

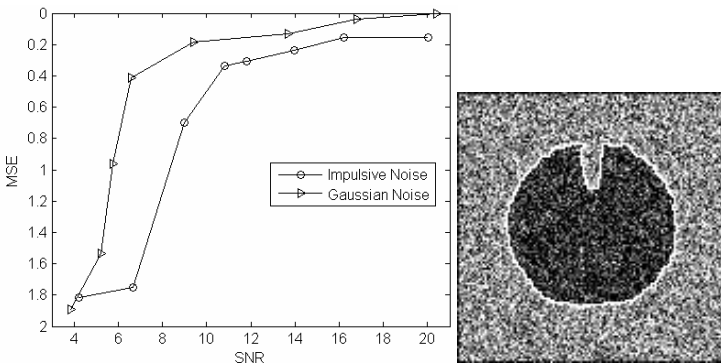


Fig. 4. Mean square error results for impulsive and Gaussian noise in different SNR levels. ($\alpha=0.5$, $0.5 \leq \lambda \leq 1$, $k=5$); and an example for Gaussian noisy image (SNR=6.58, MSE=0.41). (note that the y-axis has been reversed for the conventional curve indication purpose).

3.2 Natural Images

Natural images with complex topology are also assessed. Figure 5 shows the result for the image of a river delta with different parameters, where the river is the target object. It is suited to performance evaluation since gaps and “weak” edges exist in the image. One example is the upper part of the river, where boundaries are blurred and irregular. There are also inhomogeneous areas inside the river, which are small islands and have lower intensity. Our water flow based operator can overcome these problems. As shown in figure 5 (a), a reasonably accurate and detailed contour of the river is extracted. At the upper area, some very weak boundaries are also detected. This is achieved by using high value of k which makes the operator highly sensitive to edge response. The contour is relatively smooth by virtue of surface tension. The fluidity leading to topological adaptability is shown well by successful flow to the branches at the lower area. Most of them are detected except failure at several narrow branches. The barriers are caused either by natural irregularities inside them or noise.

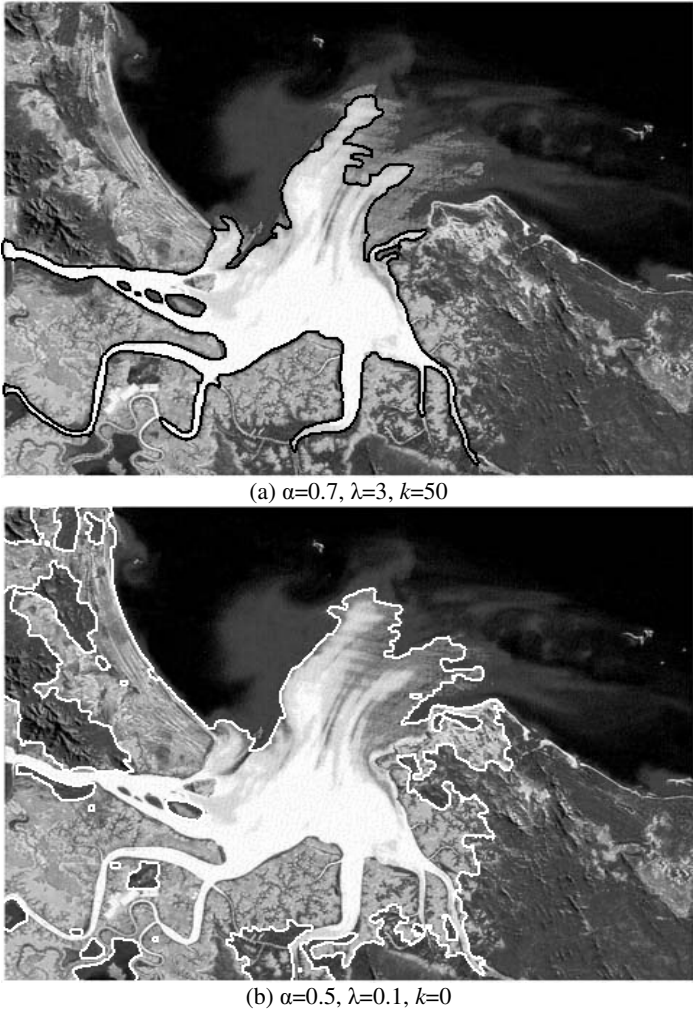


Fig. 5. Water-flow detection results for delta map image with different parameters: decreased α and λ reduce the significance of edges, and smaller k makes flow less sensitive to edges, therefore the detail detection level is lower in (b)

Different initializations inside the river were tried and with the same parameters chosen, the results are almost the same, as expected. The operator is insensitive to the source positions. By changing the parameters, however, some alternative results can be achieved. For example, figure 5(b) shows a segmentation of the whole basin of the river. It is analogy to a flood from the river. The water floods the original channels and stops at the relatively high regions. This shows the possibility of achieving different level of detail just by altering some parameters.

4 Conclusions

This paper introduces a new general feature extraction framework. The operator successfully implements the key attributes of water flow process: the fluidity, the surface tension and the adhesion. The resistance given by images is defined by a combination of object boundary and regional information. The problems of complex topological changes are solved whilst the attractive properties of snakes such as the smooth contour is retained. Those are approved by the results on both synthetic and real images. Good noise immunity is also justified both qualitatively and quantitatively. Besides, the complexity of the algorithm is relatively low. Therefore the method is expected to be of potential use in practical areas like medical imaging and remote sensing where target objects are often complicated shapes corrupted by noise.

References

- [1] L. D. Cohen, "On active contour models and balloons," *CVGIP, Image Understanding*, 53(2): 211-218, 1991.
- [2] L. D. Cohen and I. Cohen, "Finite element methods for active models and balloons for 2-D and 3-D images," *IEEE Trans. PAMI*, 15: 1131-1147, 1993.
- [3] C. Xu and J. L. Prince, "Snakes, shapes, and gradient vector flow," *IEEE Trans. Image Processing*, 7(3): 359-369, 1998.
- [4] M. Figueiredo and J. Leita, "Bayesian estimation of ventricular contours in angiographic images," *IEEE Trans. Medical Imaging*, 11: 416-429, 1992.
- [5] T. F. Chan and L. A. Vese, "Active contours without edges," *IEEE Trans. Image Processing*, 10: 266-276, 2001.
- [6] T. McInerney and D. Terzopoulos, "Topologically adaptive snakes," *Int'l Conf. Computer Vision 95*, pp. 840-845, 1995.
- [7] V. Casselles, R. Kimmel, and G. Spiro, "Geodesic active contours," *International Journal of Computer Vision*, 22(1):61-79, 1997.
- [8] R. Malladi et al., "Shape modeling with front propagation: A level set approach," *IEEE Trans. PAMI*, 17: 158-174.
- [9] R. Adams, and L. Bischof, "Seeded region growing," *IEEE Trans. PAMI*, 16(6): 641-647.
- [10] S.C.Zhu and A.Yuille, "Region competition: unifying snakes, region growing, and Bayes/MDL for multi-band image segmentation," *IEEE Trans. PAMI*, 18(9): 884-900.
- [11] V. Kiran, P. K. Bora, "Watersnake: integrating the watershed and the active contour algorithms," *TENCON 2003. Conference on Convergent Technologies for Asia-Pacific Region*, vol. 2, pp. 868-871, Oct. 2003
- [12] Bleau and L. J. Leon, "Watershed-base segmentation and region merging," *Computer Vision and Image Understanding* vol. 77, pp317-370, 2000.

Appendix

The sums of integrations in equation (10) first need be modified to discrete form:

$$F_1(C) + F_2(C) = \sum_{j=1}^{n_{int}-1} (u_j - \mu_{int0})^2 + \sum_{j=1}^{n_{ext}-1} (u_j - \mu_{ext0})^2 \quad (13)$$

where subscript “0” means the state before movement, and u_{int} and u_{ext} represent pixels inside and outside the water region respectively. After a single pixel flow, the numbers become $n_{int}+1$ and $n_{ext}-1$, respectively with corresponding changes in the statistics. By denoting the flooded pixel as u_n , we can deduce the changes. For the external term, the new term is

$$\begin{aligned}
 F_2'(C) &= \sum_{j=1}^{n_{ext}-1} (u_j - \mu_{ext1})^2 \\
 &= \sum_{j=1}^{n_{ext}-1} \left(u_j - \frac{\mu_{ext0} \times n_{ext} - u_n}{n_{ext} - 1} \right)^2 \\
 &= \sum_{j=1}^{n_{ext}-1} \left(u_j - \frac{\mu_{ext0} \times (n_{ext} - 1) + \mu_{ext0} - u_n}{n_{ext} - 1} \right)^2 \\
 &= \sum_{j=1}^{n_{ext}-1} \left(u_j - \mu_{ext0} - \frac{\mu_{ext0} - u_n}{n_{ext} - 1} \right)^2 \\
 &= \sum_{j=1}^{n_{ext}-1} (u_j - \mu_{ext0})^2 + (n_{ext} - 1) \left(\frac{\mu_{ext0} - u_n}{n_{ext} - 1} \right)^2 + 2 \left(\frac{u_n - \mu_{ext0}}{n_{ext} - 1} \right) \sum_{j=1}^{n_{ext}-1} (u_j - \mu_{ext0})
 \end{aligned} \tag{14}$$

Denote $(u_n - \mu_{ext0})$ as Δ , then $\sum_{j=1}^{n_{ext}-1} (u_j - \mu_{ext0}) = -\Delta$ (as $\sum_{j=1}^{n_{ext}} (u_j - \mu_{ext0}) = 0$), hence

$$\begin{aligned}
 F_2'(C) &= \sum_{j=1}^{n_{ext}-1} (u_j - \mu_{ext0})^2 + \frac{\Delta^2}{n_{ext} - 1} + 2 \frac{\Delta}{n_{ext} - 1} (-\Delta) \\
 &= \sum_{j=1}^{n_{ext}} (u_j - \mu_{ext0})^2 - \Delta^2 - \Delta^2 / (n_{ext} - 1) \\
 &= \sum_{j=1}^{n_{ext}} (u_j - \mu_{ext0})^2 - \Delta^2 n_{ext} / (n_{ext} - 1)
 \end{aligned} \tag{15}$$

The change to the external region is then $[-\Delta^2 n_{ext} / (n_{ext} - 1)]$. The coefficient is greater than 1, but for Δ , we have:

$$u_n - \mu_{ext0} = u_n - \frac{u_n + \sum_{j=1}^{n_{ext}-1} u_j}{n_{ext}} = \frac{n_{ext} - 1}{n_{ext}} u_n - \frac{\sum_{j=1}^{n_{ext}-1} u_j}{n_{ext}} \tag{16}$$

So the maximum of Δ is achieved when $u_n=1$ (normalized) and $u_j=0$ for others. The external change should satisfy:

$$C_{ext} = \frac{n_{ext} - 1}{n_{ext}} (u_n - \mu_{ext0})^2 \leq \frac{n_{ext}}{n_{ext} - 1} \left(\frac{n_{ext} - 1}{n_{ext}} \right)^2 = \frac{n_{ext} - 1}{n_{ext}} < 1 \tag{17}$$

Similarly, we can derive the change of the internal factor caused by the movement, which is given by:

$$C_{int} = [n_{int}/(n_{int}+1)] \cdot (u_n - \mu_{int0})^2 \quad (18)$$

The sum of the two changes gives equation (10). Since the absolute values of both terms fall in the range [0 1) for normalized images, the value range of the regional force is (-1 1), which can be directly applied to the formula

Seeded Region Merging Based on Gradient Vector Flow for Image Segmentation

Yuan He, Yupin Luo, and Dongcheng Hu

Department of Automation, Tsinghua University, Beijing 100084, P.R. China
heyuan97@mails.tsinghua.edu.cn

Abstract. Human interaction is a crucial restriction of active contour model, or snakes. In this paper we propose a fully automatic algorithm based on gradient vector flow (GVF) field and watershed-based region merging. Firstly a scalar force field is constructed by minimizing an energy function from the GVF force field. From the scalar field we extract a set of seed points facilely, and get an initial segmentation without doing curve evolution. Then a Region Adjacency Graph (RAG) based region merging algorithm is applied to get the final result. Several experimental results demonstrate that this method is efficient to multiple objects segmentation, and insensitive to noises.

1 Introduction

Image segmentation is one of the most important steps in image analysis and understanding. Its main goal is to partition an image into regions with some specific homogeneous features, such as gray scale, color, texture, etc. Generally, techniques to deal with the image segmentation problem can be categorized into four types: histogram-based, edge-based, region-based and hybrid of them. Histogram-based methods assume that homogeneous objects are manifested as clusters in the feature space. Edge-based methods are based on the abrupt changes of the feature space near object boundaries, while region-based methods are based on some specific criteria of homogeneity. However, there is no perfect algorithm for all image segmentation problem, and we need to select an appropriate method according to the type of input images.

Active contour model, or snakes, was proposed as an edge-based boundary extraction method by Kass *et al.* [1] in 1987. It is essentially a deformable contour which is forced from its initial position to approach the object boundaries by iteratively minimizing an energy function. Forces which make the contour shrink or inflate are composed of an internal component to keep its smoothness and an external component to attract the contour to object boundaries. A main restriction of snakes is setting of the initial contour. In traditional snakes the initial contour must be quite close to object boundaries, or else it may converge to wrong results or converge too slowly. Several improvements have been developed to reduce its sensitivity to initial contours, such as pressure forces [2], distance potentials [3], and gradient vector flow [4]. However, human interaction is still necessary, and in some cases it is a difficult task to set a proper initial contour.

The watershed algorithm [5] is a well known automatic region-based method for image segmentation. Region boundaries are regarded as watershed lines dividing individual catchment basins which can be extracted from the gradient image. However, the original watershed algorithm usually produces an over-segmentation result with too many catchment basins. Therefore, usually a post process is employed to merge them into the true result, such as the Region Adjacency Graph based method [6].

The algorithm proposed in this paper combines the GVF field and the watershed based image merging algorithm. Firstly we construct a scalar force field to interpret the force vector field by minimizing an energy function iteratively. The scalar force field can be regarded as the gradient image in the watershed algorithm. As a result, we can extract a set of seeds from it and get an initial segmentation by using a downstream region growing process. Finally we use a RAG based region merging algorithm to get the true segmentation. Compared with the snakes, it is fully automatic, and needs not to do curve evolution. On the other hand, it is more straightforward to get the initial segmentation from initial seeds than the watershed algorithm.

2 From GVF to a Scalar Force Field

2.1 Gradient Vector Flow

In traditional snakes, gradient vectors only exist near image edges. Therefore, the initial contour must be quite close to object boundaries, or else it will converge to wrong results or converge too slowly. Gradient vector flow [4] was proposed to increase the capture range of image edges and force contours to boundary concavities. It constructs a force vector field by diffusing the original gradient vectors from near image edges to homogeneous regions, and then takes it as the external force of a snakes.

Firstly an edge map $f(x, y)$, which has larger values near image edges and smaller values in homogeneous regions, is derived from the input image $I(x, y)$. Gradient vectors of the edge map form a vector field $\vec{\nabla}(x, y) = [u(x, y), v(x, y)]^T$. Then they are diffused by minimizing the following energy function

$$E_{GVF}(u, v) = \iint \underbrace{\mu(u_x^2 + u_y^2 + v_x^2 + v_y^2)}_{\text{smoothness energy}} + \underbrace{|\nabla f|^2 |\vec{\nabla} - \nabla f|^2}_{\text{edge energy}} dx dy \quad (1)$$

where u_x, u_y, v_x, v_y are the spatial derivatives of the vector field and μ is a real positive weight parameter to control the balance between smoothness energy and edge energy.

At points near object boundaries with a large $|\nabla f|$, the edge energy is dominant, which makes the vector $\vec{\nabla}$ close to edge gradient ∇f . On the other hand, at points in homogeneous regions with a small $|\nabla f|$, the smoothness energy is dominant, which makes the vector field vary slowly. Therefore, minimizing the energy (1) can keep the vectors near image boundaries while diffusing them away to homogeneous regions.

2.2 Scalar Force Field

Vectors in the GVF field indicate the moving direction of points on the contour. At points near image edges, force vectors point to the edges in a normal direction. Contradiction of forces from two sides makes deformable contours stop moving along the edges. On the other hand, some so-called *source points* [7] whose neighbor vectors emanate from them can be regarded as the inflating center of deformable contours, since a contour which just contains a source point will inflate until stabilizing. Consequently, we can select these source points as seeds and get an initial segmentation by region growing techniques.

Several algorithms are proposed to detect source points and perform the region growing process in a vector field. In [7] the authors identify seed points by simply checking whether all their neighbor force vectors point outward, and then get an initial segmentation by using multiple snakes to process each seed point separately. However, it can not deal with some regions with parallel vectors. Moreover, it is time consuming if there are too many source points. Then in [8] the authors further proposed a region growing method without curve evolution. They extend the GVF field to be in four directions, two of which connect the diagonal neighborhoods, and then rank each point according to the corresponding force component of the neighborhoods. After selecting points in the highest rank as seeds, they process region growing along the four-dimensional force vectors iteratively. In the algorithm they must eliminate the contradictory vectors and deal with overlaps of neighbor regions. Moreover, since the rank field is discrete, generally the initial seed points cannot grow to label all the image points, and new seeds need to be added in the un-labeled regions iteratively.

In our study we use a continuous scalar force field D to interpret the GVF vector force field \vec{v} , which can simplify the seed generation and region growing process. The scalar force field is similar to a topographic surface in the watershed algorithm. The gradient vector $\nabla D(x, y)$ at each point is just in an opposite direction of the force vector $\vec{v}(x, y)$. Therefore, values in the scalar force field decrease along vectors in the GVF field. Consequently, a source point has a maximum value among its 8-connected neighborhoods since all the neighbor forces point outward, while an edge point has a value smaller than its neighborhoods in normal direction of the edge since the force vectors in the direction point to itself.

Process to construct the scalar force field is just like the converse process of the GVF field. We minimize the following energy functional

$$E(D) = \iint \lambda D^2 + |\nabla D + \hat{v}|^2 dx dy \quad (2)$$

where $\hat{v} = \vec{v}/|\vec{v}|$ is the normalized GVF field. The first term in the integrand is used to restrict the field D within a limited range around zero, while the second term is used to make the gradient ∇D close to $-\hat{v}$. The constant λ is a regularization parameter governing the tradeoff between them.

Using the calculus of variations, we can minimize $E(D)$ in Eqn. 2 by solving the following Euler equation:

$$\lambda D - (\nabla^2 D + \hat{u}_x + \hat{v}_y) = 0 \tag{3}$$

where ∇^2 is the Laplacian operator.

2.3 Numerical Implementation

Eqn. 3 can be solved by considering D as a function of time t and solving

$$D_t(x, y, t) = \lambda D(x, y, t) - (\nabla^2 D(x, y, t) + \hat{u}_x(x, y) + \hat{v}_y(x, y)) \tag{4}$$

The steady-state solution of this linear parabolic equation is the desired solution of Eqn. 3. If we simply set the spacing between pixels and the time step to 1, and take i, j and n to index x, y and t , respectively, we will get the following iterative equation in an implicit scheme.

$$D^{n+1} - D^n = \lambda D^{n+1} - (\nabla^2 D^{n+1} + \hat{u}_x + \hat{v}_y) \tag{5}$$

For a pixel (i, j) , the partial derivative can be approximated as

$$\nabla^2 D = D_{i+1,j} + D_{i-1,j} + D_{i,j+1} + D_{i,j-1} - 4D_{i,j}$$

Consequently, we get an iterative equation to the scalar force field as follows:

$$D_{i+1,j}^{n+1} + D_{i-1,j}^{n+1} + D_{i,j+1}^{n+1} + D_{i,j-1}^{n+1} - (\lambda + 3)D_{i,j}^{n+1} = D_{i,j}^n - c_{i,j} \tag{6}$$

where $c = \hat{u}_x + \hat{v}_y$ keeps constant in the convergence.

The initial condition is set to $D^0 = 0$ for the whole field, and the energy E decreases gradually by solving Eqn. 6 iteratively as time n increases. Since the total energy is correlated to the pixel number M of the image, we use an average energy $E_{avg} = E/M$ to indicate the convergence of iterations. The convergence will be reached when the average energy value varies hardly (within a threshold).

In Fig. 1 we show an example of how a scalar force field is derived from an image. The original image and the normalized GVF field are shown in Fig. 1(a-b). By setting $\lambda = 0.01$ and resolving Eqn. 5 iteratively, we get a scalar field shown in Fig. 1(c) (the field is uniformly mapped to the gray-levels from black to white). Further more, we show the scalar field with a topographic surface plotted as a 3-D mesh in Fig. 1(d). The two source points in the object area are indicated by two high peaks, and the object boundary is indicated by a series of valleys. The convergence of the average energy is illustrated in Fig. 1(e).

3 Seeded Region Merging

As mentioned above, vectors in the GVF field are interpreted by a scalar force field, and the inflating centers of deformable contours are indicated by points

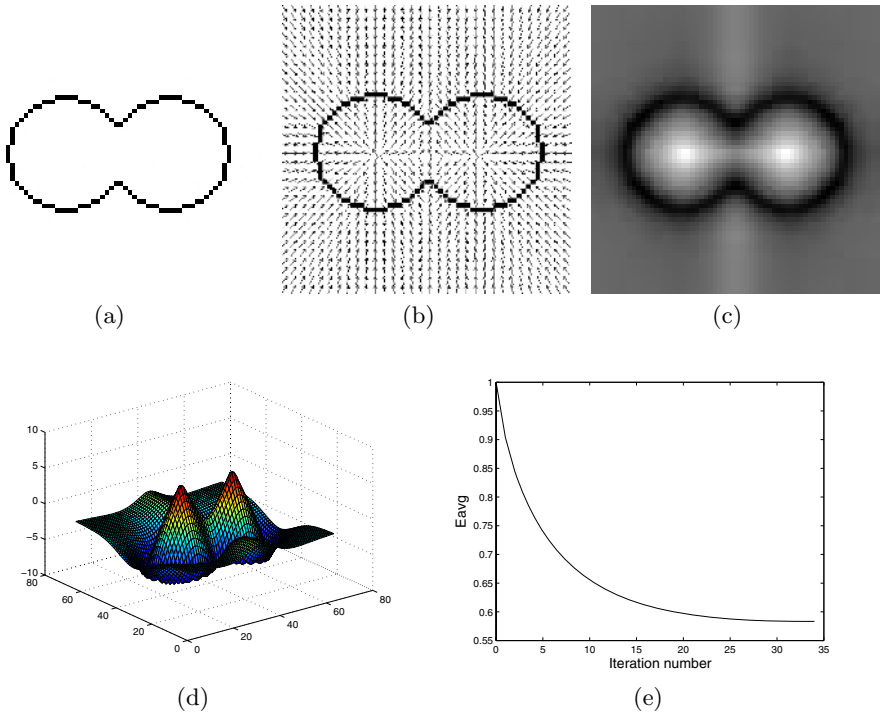


Fig. 1. From an image to a scalar force field: (a) Original image(64×64 pixels), (b) Normalized GVF field, (c) and (d) the scalar force field, (e) Convergence of average energy

in a local maximum. Therefore, we detect these centers by simply comparing their values with the 8-connected neighborhoods, and take them as seed points. Each seed grows into a region in the initial segmentation. The region growing is similar to a downstream process to simulate the curve inflation in GVF snakes. Assume that there are N seed points $S_i, i = 1, 2, \dots, N$, corresponding to N initial regions $R_i, i = 1, 2, \dots, N$, labels are simply propagated to all lower 8-connected neighborhoods. For each region R_j , we start from the corresponding seed point S_j by tracking it to label all the unlabeled lower neighborhoods, and then track these new labeled members iteratively until no more points can be labeled.

For an example of the region growing process, see Fig. 2, in which Fig. 2(a) is a scalar force field derived from a GVF field. Two seed points are selected and marked in bold. We start from the left-bottom seed point, and then the other one. The tracking sequence is illustrated by the arrows shown in Fig. 2(b). Fig. 2(c) shows the result of region growing. We note that although different orders in which seed points are processed result in different region boundaries, it is not important for the final result since the different boundaries are located in

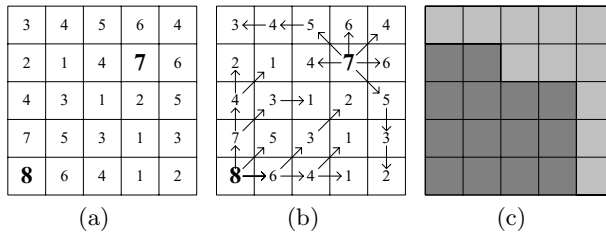


Fig. 2. An example of initial segmentation from the scalar force field. (a) A scalar force field, (b) Tracking sequence, (c) Initial segmentation.

homogeneous regions but not object boundaries. The following process of region merging will remove these differences and get a same result.

Since the object area may have more than one source point which leads to multiple regions in the initial segmentation, in our study we use the algorithm proposed in [6] to do region merging. Firstly a Region Adjacency Graph is generated from the initial segmentation, and then some adjacent regions with the minimum dissimilarity measure are merged step by step. We get the final result when no more region pair can be merged.

4 Experimental Results

In the following we give some examples of multi-object segmentation by our method.

Fig. 3 demonstrates the sensitivity of our method to noises. Fig. 3(a) is a synthetic image, and Fig. 3(b-d) are three noisy images by adding Gaussian white noise of variance 0.1, 0.3 and 0.5, respectively. It is quite difficult to set a proper initial contour for these objects if we use the snakes. Results of the four images by our method are shown in Fig. 3(e-h). The GVF field is generated by setting $\mu = 0.1$ through 50 iterations, and the scalar force field is generated by setting $\lambda = 0.01$ through 30 iterations. It is shown that as the image becomes noisy, the dissimilarity between foreground and background decreases, and object boundaries get unclear. However, our method can extract satisfied boundaries without human interaction.

More examples are illustrated in Fig. 4. For each row the three images from left to right are the original image, the initial segmentation and the final result, respectively. Fig. 4(a)(272×265 pixels) is an image of multiple cells. The GVF field is generated by setting $\mu = 0.1$ through 40 iterations, and the scalar force field is generated by setting $\lambda = 0.01$ through 30 iterations. There are 309 regions in the initial segmentation which are merged into 35 regions in the final result. Fig. 4(d)(171×200 pixels) is a magnetic resonance imaging (MRI) of a knee. The GVF field is generated by setting $\mu = 0.1$ through 20 iterations, and the scalar force field is generated by setting $\lambda = 0.01$ through 30 iterations. There are 223 regions in the initial segmentation which are merged into 7 regions in the

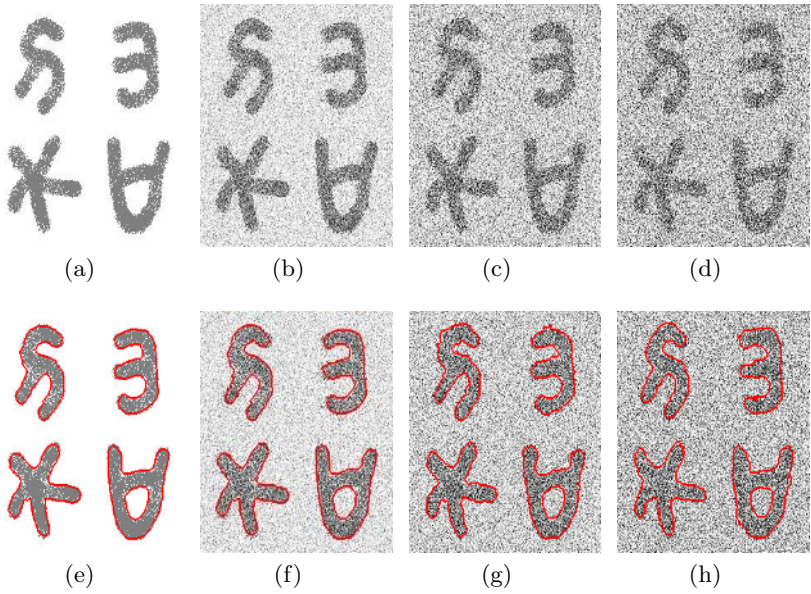


Fig. 3. Simulation of the proposed method on noisy images. (a) Original image(230×230 pixels), (b)-(d) Noisy images, (e)-(h) Results of the upper four images.

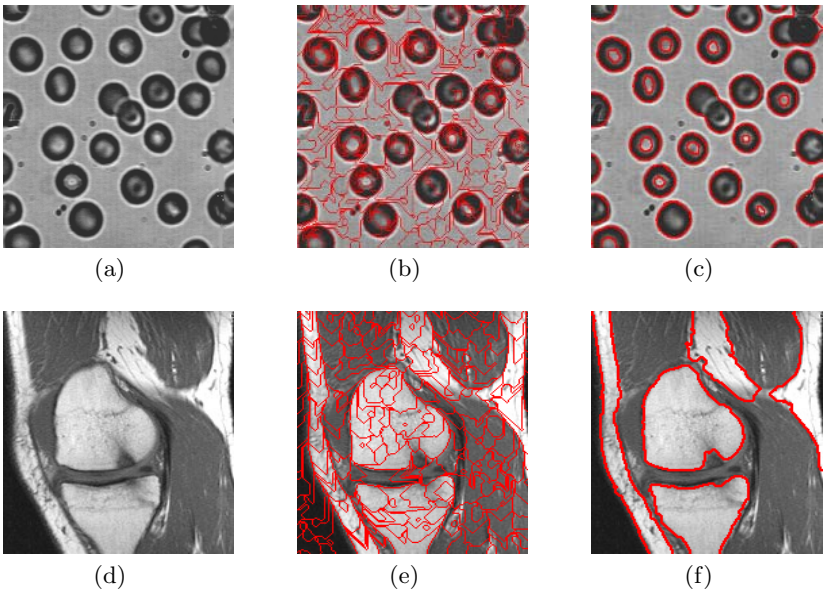


Fig. 4. Two examples of the proposed method: (a)(d) Original images, (b)(e) Initial segmentations, (c)(f) Final results

final segmentation. Although the initial segmentation in homogeneous regions seems disorderly, object boundaries are detected and identified distinctly. Region merging process can eliminate the over-segmentation and extract boundaries of multiple objects accurately and simultaneously. Moreover, the whole process is fully automatic.

We take n as the total number of pixels in an image, and m as the number of regions in the initial segmentation. Normally, m is much less than n . The time complexity of our method is composed of four components: construction of the GVF field, construction of the scalar force field, region growing, and region merging. The former two components can be transformed to the solution of an n -dimensional sparse linear equation. They both take $O(n)$ in time complexity. In the third component, selection of initial seeds needs to compare the scalar value with eight neighborhoods for each pixel, and then region growing process uses a downstream process to track unlabeled lower neighbors for each labeled pixel exactly once. Therefore, it also takes $O(n)$ in time complexity. The time complexity of region merging process is $O(mn)$ [9]. Therefore, the total time complexity of our method is $O(n + n + n + mn) \approx O(mn)$.

5 Conclusion

In this paper we present a new algorithm based on gradient vector flow for image segmentation. It needs none of human interaction, and can deal with multiple objects simultaneously. We derive a scalar force field from GVF field to avoid curve evolution. The initial centers of deformable contours in the force field are indicated with some peak points which can be detected facily. They are selected as seed points and grow into initial regions by using a simple downstream algorithm. Then a RAG based region merging process is used to overcome the over-segmentation and get the final results. The experimental results illustrate that the proposed algorithm is efficient and insensitive to noises. Moreover, compared with the snakes, it is more convenient for multi-object image segmentation.

References

1. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. *International Journal of Computer Vision* **1-4** (1987) 321–331
2. Cohen, L.D.: On active contour models and balloons. *Computer Vision, Graphics, and Image Processing: Image Understanding* **53-2** (1991) 211–218
3. Cohen, L.D., Cohen, I.: Finite-element methods for active contour models and balloons for 2-D and 3-D images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **15-11** (1993) 1131–1147
4. Xu, C., Prince, J.L.: Snakes, shapes, and gradient vector flow. *IEEE Transactions on Image Processing* **7-3** (1998) 359–369
5. Vincent, L., Soille, P.: Watersheds in digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* **13-6** (1991) 583–598

6. Haris, K., Efstratiadis, S.N., Maglaveras, N., Katsaggelos, A.K.: Hybrid image segmentation using watersheds and fast region merging. *IEEE Transactions on Image Processing* **7-12** (1998) 1684–1699
7. Yu, Z., Bajaj, C.: Normalized gradient vector diffusion and image segmentation. *Proceedings of ECCV* **3** (2004) 517–530
8. Chuang, C., Lie, W.: A downstream algorithm based on extended gradient vector flow field for object segmentation. *IEEE Transactions on Image Processing* **13-10** (2004) 1379–1392
9. Shih, F., Cheng, S.: Automatic seeded region growing for color image segmentation. *Image and Vision Computing* **23-10** (2005) 877–886

System for Reading Braille Embossed on Beverage Can Lids for Authentication

Trine Kirkhus¹, Jens T Thielemann¹, Britta Fismen¹, Henrik Schumann-Olsen¹,
Ronald Sivertsen², and Mats Carlin³

¹ SINTEF, PB 124 Blindern, N-0314 Oslo, Norway
{trk, jtt, brg, hso}@sintef.no
<http://www.sintef.no/omd>

² Tomra Systems ASA, P.O. Box 278, N-1372 Asker,
ronald.sivertsen@tomra.no

³ Carlin's Algorithm Factory, Vallerveien 152 E, N-1346 Gjøttum,
mats@carlin.no

Abstract. The paper describes a system for reading embossed Braille patterns on used aluminum beverage container lids. The intent of the system is to check whether the used containers are entitled to a refund. The lids have strong specular reflections. The reflections are avoided by a novel method that illuminates the lid alternating from two angles, and acquires two separate images. This illumination method is more compact than existing methods. We use the extended maxima algorithm to detect the Braille dots, and a cluster-based pattern point matching algorithm to recognize a pre-defined Braille pattern. The algorithms are customized to increase speed using a priori information. The system was evaluated on a test set containing 225 images. The median time used for analyzing one beverage can was 1 second, and the recognition rate was 94 percent.

1 Introduction

Recycling used beverage containers is an effective way of reducing environmental pressure. The material used in these containers – aluminum, glass or plastic – is easily reused into new containers or other goods.

Maximum environmental effect is achieved if a large proportion of the used beverage containers are recycled by the public. Deposit/refund systems effectively increase the recycled share of beverage containers. In USA the collection rate in deposit states are 72% and in non-deposit states 28% [1]. Germany recently introduced a refund system in order to achieve a higher recycled fraction of one-way containers.

There is a high labor cost in manually handling such refund systems. Therefore, various automatic machines exist that accept material for recycling and issue deposit refunds. These machines use barcode reading in combination with e.g. shape recognition to detect the container type and issue the correct refund.

Automatic machines increase efficiency at the cost of larger fraud problems. Swindlers commit fraud by tampering with containers acquired from countries

without deposit laws to make the machines accept them. Typical fraud strategies include overwriting barcodes with a corrected barcode in order to fool the machine.

To prevent fraud it is desirable to have several authentication systems. This increases the cost of fraud, and may make it uninteresting. Many such authentication systems have been developed or proposed. Shape, fluorescent tagging, RFID, printing, engraving, weight or material properties are all methods that can be used for checking the authenticity of a beverage container.

We suggest a set of embossed dots on the can lid as a new authentication method. Such dots have the benefit of being easy and cheap to include into current can manufacturing lines, unlike many of the alternatives. They are also suitable for automatic recognition by a vision system. If formed as Braille letters, such dots can have the added value of benefiting the visually impaired by carrying information about the can's content or whether the can is entitled to a refund.

This paper describes a system for verifying the presence of such dots, and recognizing one pre-defined Braille pattern. There exist numerous systems for reading Braille patterns using optical methods [2, 3, 4]. All of these have focused on patterns on paper. In this case, the dots are on highly reflective lids that require new illumination methods. Limited computing resources require optimized algorithms to facilitate recognition in real-time. The presented system is optimized to reduce cost and amount of space used.

We will first describe the chosen illumination and image acquisition system used. Then we will present the algorithms used for reading the Braille pattern and how they are optimized to facilitate the implementation on the given reverse vending machine's platform. The results of this experiment are summarized, and our conclusions are then presented.

2 System Specification

The image acquisition system had to fit into existing reverse vending machines due to backward compatibility requirements. This posed constraints on its size and geometry. The mean recognition time on the machine's platform had to be less than 1 second, and the recognition rate better than 90%. The algorithms had to be implemented on a Texas TMS320C6711 DSP processor with external SDRAM (100 MHz).

3 Illumination and Image Acquisition

Our primary focus was on designing an illumination and image acquisition system to obtain high-quality images that were easy to analyze, rather than correcting the images afterwards using time-consuming image processing techniques. We optimized the illumination to obtain easy segmentation of the can lid from the background and high contrast between dots and surface.

Beverage cans are challenging to illuminate well. Most beverage cans are produced from rolled aluminum sheets.

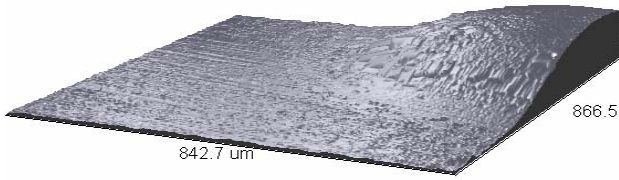


Fig. 1. An example 3D contour plot (measured with a white-light interferometer) of a rolled aluminum surface with an embossed dot of height 0.2mm and diameter 1mm in the upper right corner. There is a grating-like surface structure which makes the reflection properties direction-dependent.

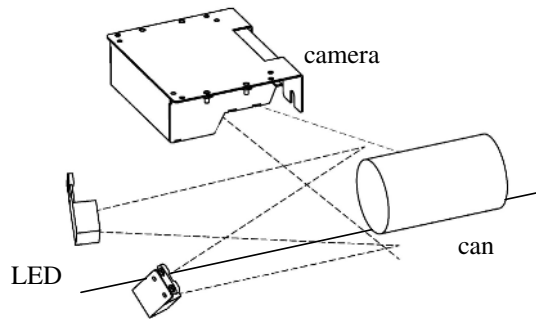


Fig. 2. Schematic drawing of the image acquisition setup. Two LED arrays (shown to the left) illuminate the can alternatively. The camera is attached to the machine's ceiling and captures images of the can lid from above.

Rolled aluminum has difficult reflectance properties due to its surface structure. The major reflection is omni directional lateral to the rolling direction and specular longitudinal to the rolling direction, due to the grating-like structure which is formed by the rolling process (fig. 1). For certain orientations of the can, a specular reflection occurs and makes the captured image unusable for further analysis.

Traditional methods avoid such reflections by using a diffuse illumination [5]. To get the illumination sufficiently diffuse, a large amount of space is required, not a possibility in this case. Only a small spot in the ceiling could be used along with some of the frontal upper interior wall.

Instead of using one large diffuse illumination source, we propose using two smaller illumination sources that can be switched on and off. Fig. 2 shows the geometric alignment of sources and camera. This idea is similar to the illumination technique used in [5], where the goal is to separate Braille (3D pattern) from an uneven background (2D pattern).

Two images are captured for each can: One image with only source one turned on, one image with only source two on. The distance between the two sources is large enough to ensure that any specular reflection does not happen in both images.

The effect can be seen in fig. 3. Fig. 3a shows an image captured with illumination source one turned on. There is a strong specular reflection in the dot region. This

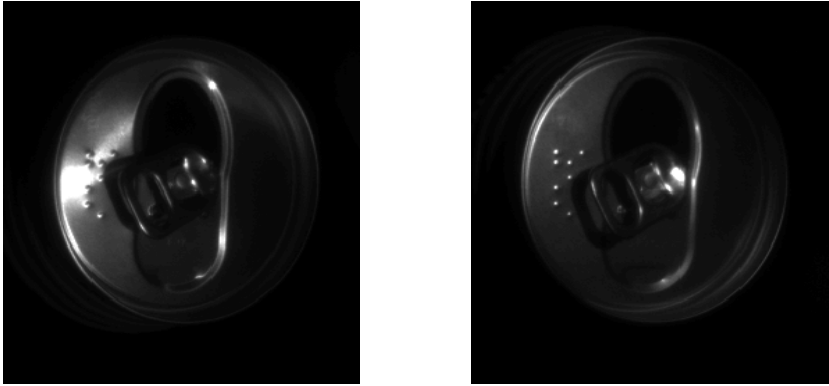


Fig. 3. Two alternative illumination sources used on the same lid. Left: Illumination source one. A strong specular reflection occurs in the dot region, making the pattern illegible. Right: Illumination source two. The reflection is not present using this illumination source.

happens due to the surface structure as described above. With illumination source two switched on (fig. 3b), the specular reflection disappears and the image can easily be analyzed [6].

3.1 Pattern Detection Methods

Cans are accepted only if they have the correct lid size and the reference dot pattern can be found on the lid. The reference pattern to be detected is shown in fig. 4. The algorithm for accepting/rejecting cans verifies this by performing three steps.

First, the lid boundary in the image is robustly detected. This makes it possible to reduce the image size significantly and estimate the perspective distortion that has occurred. The lid boundary is also used to verify that the can is of acceptable size.

Dot candidates are subsequently found. The candidates are found based on an initial thresholding followed by a subsequent dot-by-dot qualification.

Finally, the dot candidates are searched for reference pattern presence. Dot candidates are clustered and subsequently filtered to increase speed. If the pattern is found to be present, the can is accepted.

All these steps are first performed on the image captured with the left illumination source. If no match can be found here, the steps are retried on the image captured with the right illumination source. We do not attempt to detect the direct reflection first, as the frequency of direct reflexes is not large enough to warrant the extra computation time.

Due to constraints on both time and computational resources, simple and fast algorithms had to be developed.

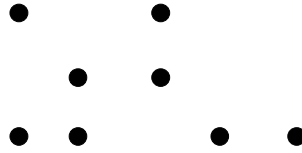


Fig. 4. Reference dot pattern. This pattern signals that the can is entitled to refund.

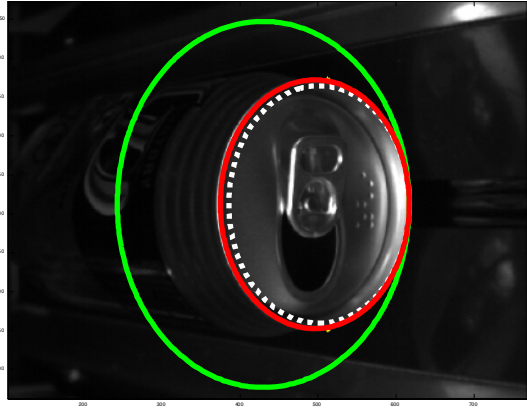


Fig. 5. The theoretical ellipses are overlaid the original image. The distance between each rim point to the ellipse is used to determine which ellipse corresponds to the imaged can.

3.2 Lid Boundary Detection

The lid is located first in order to reduce image size and verify the can size. Detecting the lid also allows us to correct the perspective distortion that is introduced by the imaging system.

The lid's location is found by first finding seven points defining the lid's rim. The detection is based on finding rapid intensity deviations from a continuously updated local background model. The points are located by using a sliding window and calculating a local threshold level based on the mean and standard deviation within that window. We search for the lid starting from the side of the image. The first significant change in pixel intensity value is assumed to be the lid's rim.

This search gives us seven points that may be along the lid's rim. After checking that the detected points are sound using per-point and inter-point constraints, we compare the points to a set of three pre-configured reference ellipses that match the three available lid sizes on the market (fig. 5). The ellipse parameters for these reference ellipses have been established using the algorithm outlined in [7]. The matching criterion is based on a Euclidean distance metric [8], measuring the distance between each of the seven points and the closest point on the reference ellipse.

3.3 Dot Candidate Detection

Possible dot candidates need to be located on the lid in order to facilitate recognition. We perform a pre-filtering using a Gaussian filter to enhance structures similar to the Braille spots before thresholding the cropped image [9].

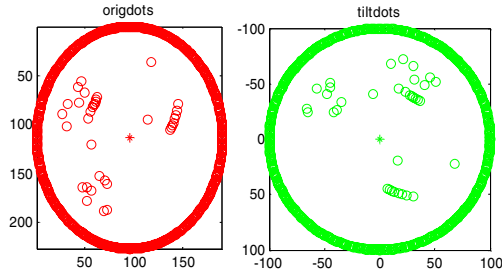


Fig. 6. Left: Detected points and detected ellipse. Right: The points are corrected according to the ellipse parameters.

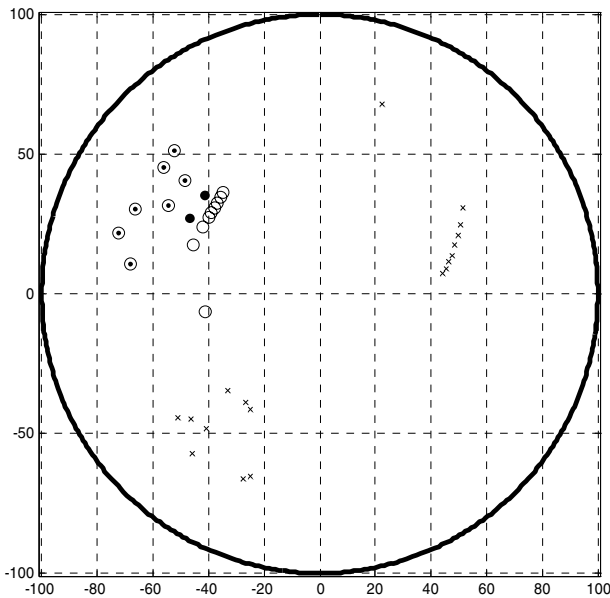


Fig. 7. Typical dot set presented to the algorithm. Crosses: Removed by the preprocessing steps. Hollow dots: Remaining for matching after preprocessing. Hollow dots containing solid dot: Dots representing the reference pattern. Solid dots: Dots in the reference pattern not found by previous steps.

The filtered image is thresholded using the extended maxima transformation [10]. This method suppresses all maxima whose depth is lower than or equal to a given threshold value h . This method is somewhat sensitive to uneven illumination, but in this case the dots are so small that the illumination locally can be assumed even.

The binary image is labeled into connected groups of pixels (blobs) [11]. The blobs are filtered using moment based blob features, and the blobs' center of gravity is used as each dot's location.

A final filtering of the blobs is done by applying an optimized version of the post-processing step of Yanowitz & Bruckstein [12]. This step removes false blobs by assuming that the average edge steepness along each blob's edge is less for false blobs than for true dots. The step is optimized by using only vertical edges, which we found gave equivalent results on less DSP time.

Before point pattern matching, the ellipse parameters detected in section 3.1 are used for compensating the perspective distortion in the blobs' center of gravity (fig. 6).

3.4 Point Pattern Matching

Point pattern matching is an old problem in computer vision, and many algorithms have been proposed. Li [13] summarizes and groups many existing algorithms. Many algorithms are unsuitable for our purposes as they are brittle (based on that angles and distances are preserved) or assume that no points are missing. In Li's taxonomy of point pattern matching algorithms, our algorithm fits into the clustering class of methods. We are able to introduce some novel optimizations due to the amount of a priori information we can use, and because we can place strict limits on the quality of acceptable pattern matches.

We have a reference pattern P containing n dots and a collection M of m dots found on the lid (the "lid dots"). We would like to test whether the pattern is present on the lid. The pattern will be assumed as present if at least $n - k$ dots can be detected, where k typically is 2 or 3.

This matching is conceptually done by selecting three dots from M , and computing an affine transformation from the selected dots. M is then transformed using the affine transform. Then we measure the maximum distance between each dot in the reference pattern and its closest dot in M (the Hausdorff distance). If this maximum distance is less than a predefined threshold for $n - k$ of the dots, the match is accepted.

Care is required to make such an algorithm fast. The number of dots found in the image will typically be quite large (fig. 7), which means that a large number of dot selections and affine transformations may have to be tested.

We have therefore developed an optimized algorithm which makes some assumptions about the placement of the pattern regarding to the lid:

- a. The pattern will have the same size on all lids.
- b. The pattern will be placed at the same distance from the lid's center on all lids.
- c. The pattern will not be rotated except from a rotation around the lid's center.

The algorithm proceeds in two steps:

1. Cluster the dots into separate clusters, and use the assumptions mentioned to throw away clusters where it can be proven that the clusters do not contain the reference pattern.
2. Match efficiently the remaining clusters against the reference pattern.

The goal of the first step is to throw away as many dots as possible. The remaining dots should be clustered into as small clusters as possible. Small clusters are desirable,

as the matching step following this part has a complexity of $O(c^2n)$ where c is the size of the cluster and n is the number of dots in the reference pattern.

We will now describe four $O(n)$ steps and one $O(n^2)$ step that reduce the amount of dots passed to step two greatly. As these steps have low computational complexity compared to step two, the complexity of the algorithm as a whole is reduced.

Dot Clustering and Removal

As the pattern has the same size, position and rotation on all lids, the maximum and minimum distance from the lid’s center to any dot within the reference pattern will be the same for all lids. This means that any dot found to be closer to the center than the minimum distance – or farther away than the maximum distance – can be thrown away.

The other filters cluster lid dots and analyze the found clusters. In order to do this efficiently, we use polar coordinates (d_i, θ_i) for the lid dots, and sort the lid points according to θ_i .

Feature Extraction from the Reference Pattern

Some features are required from the reference pattern (fig. 8). These can be extracted beforehand to gain speed. Three features are used:

- a. θ_{cmax} : The maximum angle difference between two consecutive dots given that up to k dots are missing
- b. θ_{max} : Maximum angle difference between two arbitrary dots in the reference pattern given that up to k dots are missing
- c. D_{max} : Maximum Euclidean distance between two arbitrary dots in the reference pattern given that up to k dots are missing

The features can be calculated efficiently by presorting the dots according to angle prior to calculation.

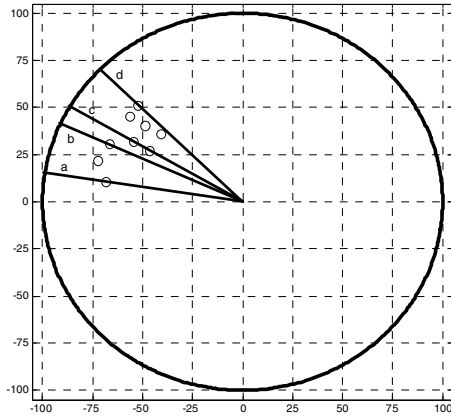


Fig. 8. Reference pattern (circles) and pattern properties used by the algorithm. Angle b-d: θ_{max} indicating the maximum angle span of the pattern given that $k=2$ dots are missing. Angle a-c: θ_{cmax} , maximum angle span between two consecutive dots given that $k = 2$ dots are missing.

Clustering Dots Using the Angular Distance between Two Dots

The lid dots are sorted based on θ , and put into one cluster. If two consecutive lid dots have an angular distance that exceeds θ_{max} , then these two lid dots cannot simultaneously be part of the reference pattern (one of them may, though). The cluster thus can be split between these two dots, forming two smaller clusters. These two clusters can be checked subsequently for further similar split dots.

The result of this operation will be that the single large clusters have been split into many smaller clusters.

Clusters that do not contain at least $n - k$ dots or span at least θ_{max} can subsequently be removed.

Removing Non-Dense Parts of Remaining Clusters

The previous operations may leave clusters that are large (thus not removable based on the θ_{max} criteria) and with sufficient number of dots. However, the dot density within parts of such a cluster may be too low to actually contain the pattern. Only those parts of the cluster where at least $n - k$ dots can be found within θ_{max} can contain the reference pattern. The remaining parts of the cluster can be ignored. This can be implemented in $O(q)$ where q is the number of dots remaining.

Removing Too Large Clusters

There is a maximum distance d_{max} between two dots in the reference pattern. If two lid dots are more than d_{max} units apart, they cannot simultaneously match the reference pattern. This may mean that they can be put into two separate clusters. This determination can be done in $O(c^2l)$, where l is the number of remaining clusters and c is the number of dots within each cluster.

Braille Pattern Recognition

After the previous steps, only one or two dot clusters remain. We now proceed to check whether at least $n - k$ dots match the reference pattern. While in our assumptions we have listed that the match should succeed only in the case of pure rotation around the lid's center, imprecision in our previous steps does not make this possible. For instance, our perspective correction algorithm makes idealized assumptions about the box's position within the machine.

We therefore have to do pattern detection allowing a complete affine transformation.

Our detection strategy is an optimized exhaustive search. We select three dots from the reference pattern and three dots from the lid dots. These dots are then used to build the affine transformation that transforms the lid dots into the coordinate space of the reference pattern. We then verify that $n - k$ dots in the reference pattern have a corresponding dot close enough in the cluster dots.

Assuming that there are d dots in the cluster, this would give a performance of $O(d^3n^3)$ as all combinations would have to be tested. Our requirement that at least $n - k$ dots need to match allows for one additional optimization. Let R be the first $k + 1$ dots in the reference pattern, and N be the remaining $n - k - 1$ dots. If at most $n - k$ dots are allowed to be missing, then at least one of the dots in R must be present in the cluster dots. It is thus sufficient to test $k + 1$ dot combinations from the reference dots. This brings the algorithm's performance to $O(n^3)$.

Table 1. Recognition performance per illumination type. A total of 225 image pairs were used for testing the algorithms.

Image	Number of successful decodes	Success %
Left image	182	80,9
Right image	195	86,7
Both images	211	93,8

Table 2. Number of cluster and dots per cluster before and after preprocessing. Mean and one standard deviation shown for all values except “max dot count”. Images where no dots were found have been represented as having zero clusters, thus creating an average cluster count less than one before preprocessing.

	Before preprocessing	After preprocessing
Cluster count	0.996±0.07	1.3±0.6
Dot count/cluster	20±7	11±5
Max dot count	52	41

In addition, we are placing strict limits on the affine transformation in order to gain speed and reduce the amount of false positives.

4 Experimental Results

4.1 Recognition Performance

The algorithms were tested using a test set containing 15 cans, which were imaged in 15 different positions and rotations each. Two images were captured for each position, one image for each illumination source. A total of 225 image pairs were used for testing the algorithms.

Table 1 summarizes the recognition performance. Using only one illumination source gives roughly 10% less decode performance than using both illumination sources and combining the results. This is due to our ability to handle specular reflexes on the can lids. In approximately 6% of the cases both decode attempts fails. The cause is the can lever opener, which blocks the camera’s view of the Braille pattern.

The system was tested with approximately 1000 cans without the embossed pattern. We found that the amount of false positives was less than 5%.

4.2 Algorithm Performance

In order to decrease the algorithm’s runtime a number of steps were performed in order to reduce the number of dots in each cluster.

Table 2 summarizes the number of clusters in each image before and after the preprocessing step is done. On average, the number of dots in each cluster is halved, which makes the second step of the algorithm significantly less time-consuming.

The time constraints were < 1 second for capturing and reading the Braille. We achieved a median of ~1 second analyzing time, but some images with many dot candidates used longer processing time.

5 Summary and Conclusion

We have described a complete system for reading embossed Braille dots on aluminum can lids. The rolled aluminum surface is highly specular and reflections will occur depending on the rolling direction versus camera and illumination geometry. Our illumination method handles this by using less space than e.g. diffuse illumination for achieving the same means.

Low cost is essential in making automatic reverse vending machines widely available. The cost requirement limits available computing resources. We have developed specific and optimized algorithms that employ a priori information to make this possible on the selected computing platform within the given time constraints (~1 second).

Refunds are one of many incentives offered to customers to begin and continue recycling. A positive experience when using the reverse vending machine is important as well. A low amount of false negatives (rejects) is critical to achieve this.

In our experiment, the system recognized the pattern in 94% of the cases. However, our experiments were performed on one reverse vending machine. The machine and the used beverage cans were clean, the external illumination was under control and the variations in cans were minimal. This artificially increases decoding performance. For a real-life system, we expect a performance of 90% recognition rate. Still, we expect this to be sufficiently high to satisfy the customer.

Fraud rates must be kept low in order to make refund systems possible. Otherwise, bottlers, and groceries will refrain from participating in the system. The described system makes refund fraud more difficult and may alleviate this concern.

References

1. SEARS Report, New York, Bottle bill study, February 2004
2. T.W. Hentzschel, P. Blenkorn, An optical reading system for embossed braille characters using a twin shadows approach, *Journal of Microcomputer Applications*, 18:4, pp 341-354, 1995.
3. N. Falcón, C. M. Travieso, J. B. Alonso, M. A. Ferrer, Image Processing Techniques for Braille Writing Recognition, *Computer Science*, Vol 3643, 2005.
4. Antonacopoulos, D. Bridson, A Robust Braille Recognition System, *Computer Science* Vol 3163, 2004.
5. B.G. Batchelor, Lighting and viewing techniques. In: Batchelor, B.G., Hill, D.A. & Hodgson, D.C (eds) *Automated Visual Inspection*, IFS Publications, North-Holland, 1985.
6. R. Sivertsen, M. Carlin, B.G. Fismen, I-R. Johansen, Device for recognising containers, WO 2004/003830, September 15, 2005.
7. R. Halíř, J. Flusser, Numerically Stable Direct Least Squares Fitting of Ellipses, *Proc of the 6th Int'l Conf in Central Europe on Computer Graphics and Visualization*, WSCG '98, pp. 125-132, 1998.
8. E.R. Davis, *Machine Vision: Theory, Algorithms, Practicalities*, pp. 269-271, Academic Press, 1990.
9. W. K. Pratt, *Digital Image Processing*, 2nd edition, John Wiley & Sons Inc., NY 1991.

10. P. Sollie, *Morphological Image Analysis: Principles and Applications*, 2nd edition, Chap. 6.3.4, Springer Verlag, 2003.
11. R.M. Haralick, L.G. Shapiro, *Computer and Robot Vision, Volume I*, Addison-Wesley, 1992, pp. 28-48.
12. S.D. Yanowitz, A.M. Bruckstein, A new method for image segmentation, 9th Int'l Conf on Pattern Recognition, vol 1, pp 270-275, 1988.
13. B. Li, Q. Meng, H. Holstein, Point pattern matching and applications – a review, in *Proc. Int'l Conf. Systems, Man and Cybernetics*, 2003, pp. 729–736

Leukocyte Segmentation in Blood Smear Images Using Region-Based Active Contours

Seongeun Eom, Seungjun Kim, Vladimir Shin, and Byungha Ahn

Department of Mechatronics,
Gwangju Institute of Science and Technology,
1 Oryong-dong, Buk-gu, Gwangju 500-712, South Korea
{seueom, zizone, vishin, bayhay}@gist.ac.kr

Abstract. In this paper, we propose a segmentation method for an automated differential counter using image analysis. The segmentation here is to extract leukocytes (white blood cells) and separate its constituents, nucleus and cytoplasm, in blood smear images. For this purpose, a region-based active contour model is used where region information is estimated using a statistical analysis. The role of the regional statistics is mainly to attract evolving contours toward the boundaries of leukocytes, avoiding problems with initialization. And contour deformation near to the boundaries is constrained by an additional regularizer. The active contour model is implemented using a level set method and validated with a leukocyte image database.

1 Introduction

A leukocyte (white blood cell) differential count as a percentage is one of the most frequently performed blood tests and plays an important role in the diagnosis of diseases such as anemia. In hospital, manual differential is usually performed by taking a drop of blood, spreading it on a slide, staining it, and evaluating around 100 cells for quantity and quality. However, it is tedious and time consuming to locate, classify and count leukocytes. An automated differential counter using image analysis makes it possible to replace the work, reducing reporting time and increasing precision with the larger of number of cell counted. The counter system is normally performed in this procedure: localization, segmentation, feature extraction and classification. We here deal with the leukocyte segmentation problem which is most difficult and error-prone.

In an attempt to solve the problem, several approaches have been presented in the literature. Wermser et al. [1] and Cseke [2] use hierarchical thresholding with two color features which discriminate between constituents of blood smear images. A more sophisticated algorithm is proposed by Sinha and Ramakrishnan [3]. It estimates a mixed Gaussian model using Expectation-Maximization (EM) algorithm in HSV color space, and thresholds and labels each Gaussian component with a prior knowledge. It is obvious that the algorithm is able to improve performance over linear thresholding, but it still does not use spatial information and so is limited to segment leukocytes.

To overcome the limitation, another approach using region information is proposed. Haussman et al. [4] uses region labeling by relaxation operation. Initial regions are obtained from split-merge method and then relaxed using relationship between adjacent regions. [5] is also a similar method based on fuzzy rules. Although this approach uses both local and global information, it is not straightforward to design rules (or relationship) covering various situations.

More recent literature has applied active contour models since it is useful to detect objects in an image by evolving contours. One of the advantages of the model is that contours can be controlled by its geometric properties and external properties from an input image in a single framework. Furthermore higher level information such as shape priors also can be incorporated into the same framework. Hence we can design an active contour model using information suitable for a specific application.

The active contours proposed in [6,7,8] are based on the gradient of an input image. However a high gradient is often shown at the cytoplasm of polymorphonuclear cells (granulocytes). It means that evolving contours are likely to be stuck at local minima and so the performance is heavily influenced by initial contours. [9] instead uses the gradient vector flow (GVF) calculated from boundaries of cells including erythrocytes (red blood cells). The method relieves the initialization problem but contours still could be stuck in the area of small GVF force or attracted to one side of the boundary depending on initial contours

In this paper, we propose a more robust method using a region-based active contour model. Region information here is estimated through statistical analysis on intensity features of input images. Thanks to the regional statistics, initial contours are placed safely on each leukocyte and propagate toward its boundary, avoiding the local minima problem. At earlier stage, evolving contours are mainly under influence of the region statistics, but their deformation is controlled by an additional regularizer as they come close to the boundaries of leukocytes. The regularizer contributes to forming the shape of a leukocyte

2 Proposed Method

The proposed method is composed of five steps as shown in Fig. 1. We first select significant intensity features in blood smear images and estimate a finite mixed Gaussian model using the EM algorithm where the number of components and initial parameters are chosen with intensity priors. With the estimation, Bayes probabilities of each constituent of a leukocyte are computed on piecewise constant partition and then are embedded into an active contour model as region information. In the following subsections the procedure is explained in detail.

2.1 Regional Statistics

Each constitute of blood smear images shows the characteristics discriminating one another such as color, texture, and shape. But it seems that color information is the most significant. Depending on color model, various combination of

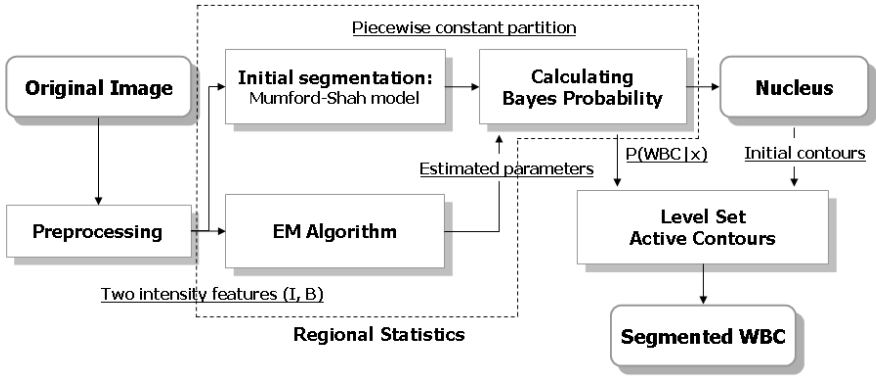


Fig. 1. Block diagram of the proposed segmentation method for leukocyte images. It is based on region-based active contours where the region information is estimated using a statistical analysis.

channels can be used. The experiments shows that among them grayscale and blue intensity space (I, B) is the most promising. Fig. 2 shows the typical pattern of the features extracted from a smoothed image with a Gaussian filter, and also exhibits the main location of each constituent.

As can be seen in the feature pattern, it is difficult to discriminate between different components through simple thresholding or clustering approach. For this reason, we estimate parameters of a finite mixed Gaussian model with the Expectation-Maximization (EM) algorithm, assuming all the components have a Gaussian distribution.

The EM algorithm [10] consists of two steps, Expectation and Maximization. The first step is to calculate the posterior probabilities with initial parameters for each data point, and then update the means, the covariance matrices and the mixing coefficients for each component in the second step. The procedure is repeated until a variation is less than some fixed threshold.

An initial guess, however, should be carefully determined particularly for blood smear images, because there often exit one or two more Gaussian components in the feature space: for example, at the granule in cytoplasm or outer boundary of an erythrocyte. At first we could assume five or six components, but it does not work well in two reasons. (1) Additional components are likely to have a negative effect on the estimation of the others, and (2) even not so it is not straightforward to label to four partitions.

Therefore the number of components is set to be four and now a way of preventing components from converging to unexpected local minima is required. The K-means algorithm [10] are commonly used to select initial parameters. However, it is not appropriate in our case for the same reason mentioned earlier: existence of undesirable Gaussian components. Instead those are determined based on the characteristics of the feature pattern and intensity priors of each constituent. Left-lower (A) and right-upper (B) feature point are selected as

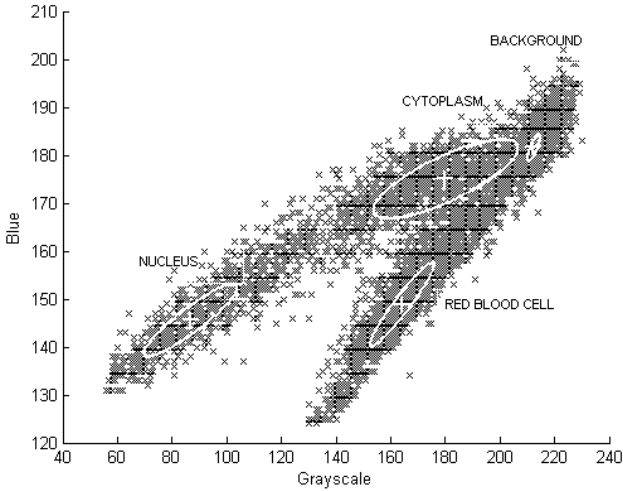


Fig. 2. Typical example of the feature pattern and the main location of each constituent: nucleus, cytoplasm, red blood cell, and background

an initial guess of nucleus and background respectively. And an erythrocyte takes a point (C) located the farthest from the line going through two points, A and B. the last one (D) is decided middle point of two points, A and C. This simple method has shown more better results than the K-means algorithm. Furthermore, labeling is easily done by the same way with the final mean points

With the estimated and labeled Gaussian components, the posterior probabilities for each constituent can be calculated using Bayes' theorem [10]:

$$P(\omega_j|x) = \frac{p(x|\omega_j)P(\omega_j)}{p(x)},$$

where

$$p(x) = \sum_{j=1}^4 p(x|\omega_j)P(\omega_j)$$

and ω_j is a j th class, namely, nucleus, cytoplasm, erythrocyte or background, and x is a feature vector (I, B). It could be applied to each pixel itself, but we here compute the Bayes probabilities on piecewise constant partition which is obtained by minimizing the Mumford-Shah functional [11]:

$$F_1(u, B) = \int_{R-B} (u - u_0)^2 + \lambda \text{length}(B),$$

where u_0 is an input image defined on a set R and u and B are a piecewise constant function and the boundaries between regions, respectively. This is useful to avoid many isolated points and small holes, and form more accurate boundaries.

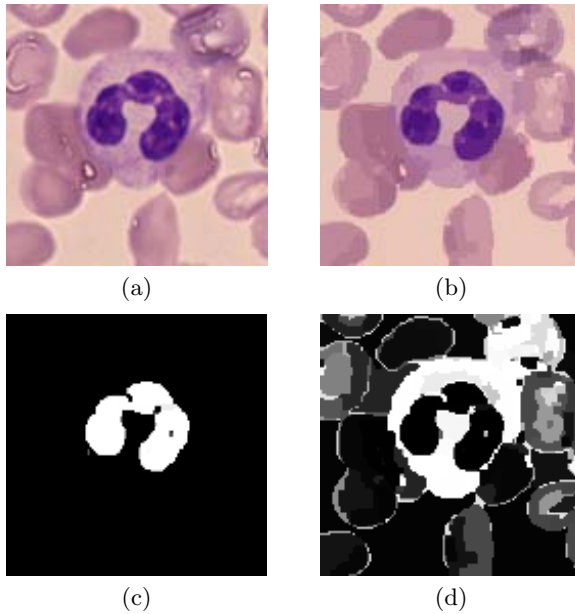


Fig. 3. Regional statistics. (a) input image, (b) piecewise constant partition, (c-d) posterior probability map of nucleus and cytoplasm. Courtesy of CellAtlas, CellaVision AB [12] (with the following figures).

Fig. 3 shows an example of initial partition and probability map of a leukocyte. The probability map is used in the region-based active contours framework as region information in the next stage.

2.2 Active Contours for Leukocyte Segmentation

Active contour models [13], also called snakes, are a powerful segmentation tool in a variety of image processing applications. Many models have been developed based on their representation (explicit or implicit) and information (edge and/or region) used. Hence, from the point of view of application the selection of an appropriate model is very important.

In our case the following implicit model is applied which is based on region information computed from the intensity features:

$$\frac{\partial \phi}{\partial t} = \delta(\phi)[\alpha_1(P - C_1)^2 + \alpha_2(P - C_2)^2] + \beta\kappa|\nabla\phi| + \gamma\mathbf{V} \cdot \nabla\phi,$$

where ϕ is a implicit level set function [14,15] where the boundary (interface) is defined by $\phi = 0$, and α_1 , α_2 , β , and γ are positive parameters. An input image P , here, is the posterior probability of a leukocyte $P(\text{WBC}|x)$ computed earlier and C_1 , C_2 are the averages of P inside and outside of propagating contours. After all the first two terms [16] attract contours toward the boundaries of a

leukocyte. As shown in Fig. 3, however, part of neighboring erythrocytes also could have a high probability and so regularization is needed.

A leukocyte is normally close to the shape of a circle but is likely to deform. In this case, however, the deformation is “smooth”, while undesirable regions having a high probability tend to be “rough” as shown in Fig. 3. Therefore a regularization method for discriminating the two cases is required. First a parameter of mean curvature term $\kappa|\nabla\phi|$ is set to be high so that the shape is close to a circle, and then a deformation beyond that is driven by the gradient vector flow (GVF) $\mathbf{V}(u, v)$ [17] obtained by minimizing

$$F_2(\mathbf{V}) = \int \int \mu(u_x^2 + u_y^2 + v_x^2 + v_y^2) + |\nabla f|^2 |\mathbf{V} - \nabla f|^2 dx dy,$$

where an input image f is the boundaries of regions Bayes classified as a leukocyte. In spite of constraint of curvature, the contour now tries to propagate toward smoothly deformed boundaries thanks to this externally generated velocity field. On the other hand, it does not try in roughly deformed regions because the normal force of GVF is very weak at the boundary between a leukocyte and an erythrocyte [9].

Besides this role, GVF contributes to extracting the contours with the first two terms in the initial stages and so seems to be able to take the place of the first two ones. However, in this case part of the contours could be stuck in the area of small GVF force. Gradient information also causes the similar stuck problem at the local minima and so is not used in the proposed model.

Another important factor in the use of active contours is initial contours. In our case it is relatively insensitive to the initial contours which just need to be located inside leukocytes. Fortunately Bayes classified nuclei are satisfactory in the regional statistics stage. Therefore nuclei regions are extracted at the stage and are used for selection of initial contours. We put circles at the centroid of each nucleus, whose diameters are 0.3 times ones of circles with the same area as each nucleus. If there are multiple nuclei in a leukocyte they are merged automatically thanks to level set implementation. Fig. 4 shows the propagating contours on the posterior probability map.

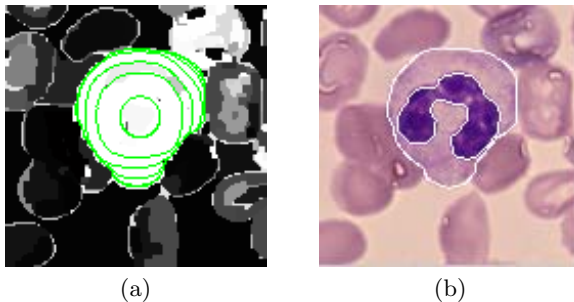


Fig. 4. Segmentation using region-based active contours. (a) propagating contour, (b) segmentation result.

3 Experimental Results

The experiments were carried out on the cell image database which was provided by CellAtlas [12]. All the images in the database were stained by the May-Grünwald-Giemsa (MGG) method and already classified by type by experts within the field of hematology. The types of leukocytes used in our experiments are five: neutrophil, lymphocyte, monocyte, eosinophil, and basophil. Assuming that leukocytes were localized, a region of interest (ROI) containing leukocytes was extracted with 128×128 size.

These images first were smoothed with a 3×3 Gaussian filter with a standard deviation of 0.8 and in the Mumford-Shah functional for a piecewise constant partition, a region-growing method [18] was applied, where the number of final regions (1000) was used as a stopping criterion instead of a scale λ . In the last stage, there are three important parameters which were set ($\alpha_1 = \alpha_2 = 0.005, \beta = 4.5, \gamma = 0.6$) as follows: (1) with $\gamma = 0$, adjust the first two parameters such that a contour becomes similar to a circle in shape as it approaches to the boundary of a leukocyte and then (2) γ value is chosen for the GVF force to push the contour toward the smoothly deformed regions.

Fig. 5 shows the process of segmentation applying the proposed method under the conditions above. The first column is the original image for each type of leukocytes, and the second shows evolving contours on the Bayes probability map of a leukocyte. The final segmentation results are shown in the last column. As already expected, errors happen in parts of images in Fig. 5(b), but in the most of cases the errors happen between cytoplasm and red blood cells because of similarity of their intensity features (I, B). However, it is noteworthy that the probability of errors in cytoplasm regions is very low and so true leukocytes are mostly included in the regions having high probability $P(\text{WBC}|x)$ like in Fig. 5(b). The used features (I, B) were selected among two channels or full channels of RGB, HSV, $L^*a^*b^*$, and $L^*u^*v^*$ space, not scaled, using the probability of error in the Bayes classification. More sophisticated methods could be used for optimal feature extraction, but it is now acceptable.

As explained earlier Bayes classified nuclei regions are satisfactory but sometimes there are undesirable small segments and holes and so we applied the morphological operators, area opening and closing, in order to remove the undesirable regions. The postprocessed nuclei were then used for initial contours. If there are multiple nuclei the same number of initial contours are placed as shown in Fig. 5(b). At an earlier stage these contours propagate mainly under influence of the regional statistics, and deform under influence of the curvature and GVF force as they come close to the boundary of a leukocyte. Fig. 5(b) shows the process of propagation. It exhibits that a contour is able to cross small holes inside a leukocyte and multiple ones are finally merged.

To validate the segmentation results shown in Fig. 5(c), we adopted the most common approach. The boundaries of leukocytes were first extracted manually and compared to the obtained results. The following metric [19] was used as a measure of segmentation accuracy: $S = 2 \cdot n\{R \cap T\} / (n\{R\} + n\{T\})$. This measure quantifies similarity $S \in [0, 1]$ between two segmentations R and T .

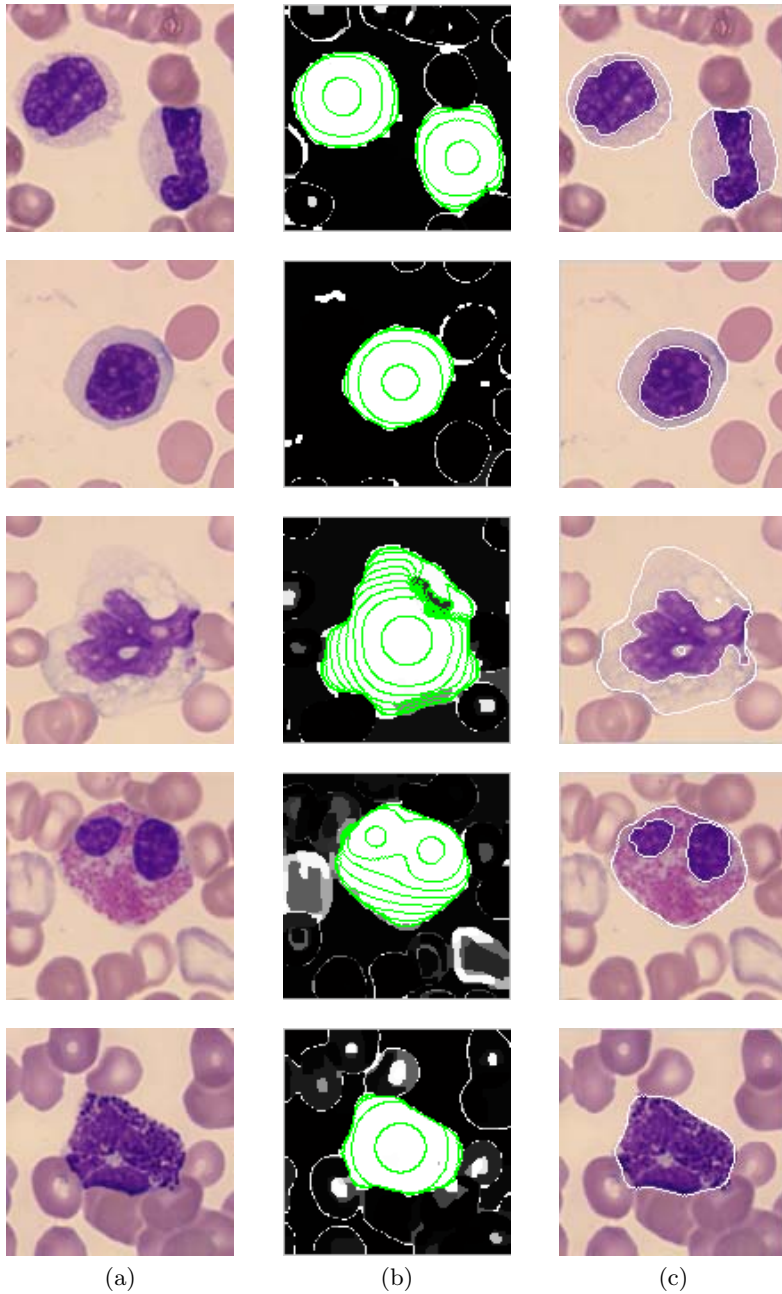


Fig. 5. Process of segmentation using the proposed method. (a) original images: neutrophil, lymphocyte, monocyte, eosinophil, and basophil, (b) intermediate results on the Bayes probability of a leukocyte, (c) final segmentation results.

Table 1. Average segmentation accuracy rates and their standard deviations (%)

Cell type	Neutrophil	Lymphocyte	Monocyte	Eosinophil	Basophil
# of images	80	125	120	124	154
Nucleus	91.40 (13.94)	95.45 (2.94)	90.83 (11.03)	88.43 (15.26)	- (-)
Cytoplasm	93.60 (6.85)	95.10 (4.45)	91.82 (10.24)	91.41 (6.91)	93.02 (9.92)

$n\{R\}$ indicates the area of R . Table 1 provides the average segmentation accuracy rates and their standard deviations of nucleus and cytoplasm for all the images; the nucleus of basophil was not included in the evaluation because it is often obscured by cytoplasmic granules and so is difficult to discern. We have obtained the reasonable results with an overall average error less than 8%.

The major errors were caused by the following reasons. First, when touching regions of erythrocytes have a high probability $P(\text{wBC}|x)$ and the regions look like part of a true leukocyte, no distinction is made between them and so a contour propagates into the regions. Second, when cytoplasmic granules are as dark as a nucleus or a cytoplasm as bright as a background, they are classified as a nucleus and a background, respectively. Third, when the shape of a leukocyte is deformed excessively, a contour fails to deform that much. The parameters involving deformation was selected as tradeoff between this case and the first one. In future research we should focus on these problems in order to improve performance more accurately.

A natural extension of the proposed method is to deal with touching leukocytes using a single level set representation. It is obvious that the separate leukocytes one another can be segmented well and also even touching ones can be segmented individually, but not at a time in a single level set framework. So we are currently under investigation on this point and expect to be useful for other applications, too.

4 Conclusion

We have proposed a segmentation method of leukocytes in blood smear images. It is based on a region-based active contour model driving initial contours toward the boundary of a leukocyte, avoiding problem with initialization and local minima. Region information here is estimated from intensity features which discriminate effectively between the constituents of blood smear images. And a regularizer has also applied in the same model to constrain excessive deformation of a evolving contour. In the experiments with a public image database, we have obtained the segmentation results with an overall average error less than 8%. Although it can be more improved by decreasing the errors mentioned earlier, the results are reasonable as an input for leukocytes classification.

References

1. Wermser, D., Haussmann, G., Liedtke, C.E.: Segmentation of blood smears by hierarchical thresholding. *Computer Vision, Graphics, and Image Processing* **25** (1984) 151–168
2. Cseke, I.: A fast segmentation scheme for white blood cell images. In: *Proc. 11th IAPR Int. Conf. Pattern Recognition, Conf. C: Image, Speech and Signal Analysis. Volume 3.* (1992) 530–533
3. Sinha, N., Ramakrishnan, A.G.: Automation of differential blood count. In: *Proc. Conf. Convergent Technologies for Asia-Pacific Region TENCON 2003. Volume 2.* (2003) 547–551
4. Haussmann, G., Liedtke, C.E.: A region extraction approach to blood smear segmentation. *Computer Vision, Graphics, and Image Processing* **25** (1984) 133–150
5. Park, J., Keller, J.M.: Fuzzy patch label relaxation in bone marrow cell segmentation. In: *Proc. IEEE Int. Conf. Systems, Man, and Cybernetics.* (1997) 1133–1138
6. Ongun, G., Halici, U., Leblebicioglu, K., Atalay, V., Beksac, M., Beksac, S.: An automated differential blood count system. In: *Proc. 23rd EMBS Int. Conf.* (2001) 2583–2586
7. Nilsson, B., Heyden, A.: Segmentation of dense leukocyte clusters. In: *Proc. IEEE Workshop on Mathematical Methods in Biomedical Image Analysis.* (2001) 221–227
8. Nilsson, B., Heyden, A.: Model-based segmentation of leukocytes clusters. In: *Proc. 16th International Conf. Pattern Recognition. Volume 1.* (2002) 727–730
9. Theerapattanakul, J., Plodpai, J., Pintavirooj, C.: An efficient method for segmentation step of automated white blood cell classifications. In: *Proc. IEEE Region 10 Conf. TENCON 2004. Volume 1.* (2004) 191–194
10. Duda, R.O., Hart, P.E., Stork, D.G.: *Pattern Classification.* 2nd edn. Wiley-Interscience, New York (2000)
11. Mumford, D., Shah, J.: Optimal approximations by piecewise smooth functions and associated variational problems. *Communications on Pure and Applied Mathematics* **42** (1989) 577–684
12. (<http://www.cellatlas.com>)
13. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. *Int. J. Computer Vision* **1** (1988) 321–331
14. Osher, S., Fedkiw, R.: *Level Set Methods and Dynamic Implicit Surfaces.* 1st edn. Springer, New York (2002)
15. Sethian, J.A.: *Level Set Methods and Fast Marching Methods.* 2nd edn. Cambridge University Press, Cambridge (1999)
16. Chan, T.F., Vese, L.A.: Active contours without edges. *IEEE Trans. Image Process.* **10(2)** (2001) 266–277
17. Xu, C., Prince, J.L.: Snakes, shapes, and gradient vector flow. *IEEE Trans. Image Process.* **7(3)** (1998) 359–369
18. Aubert, G., Kornprobst, P.: *Mathematical Problems in Image Processing. Volume 147 of Applied Mathematical Sciences.* Springer, New York (2001)
19. Zijdenbos, A.P., Dawant, B.M., Margolin, R.A., Palmer, A.C.: Morphometric analysis of white matter lesions in mr images: method and validation. *IEEE Trans. Medical Imaging* **13(4)** (1994) 716–724

Multiresolution Lossy-to-Lossless Coding of MRI Objects

Habibollah Danyali¹ and Alfred Mertins²

¹ Department of Electrical Engineering
University of Kurdistan, Sanandaj, Iran
hdanyali@ieee.org

² Signal Processing Group, Institute of Physics
University of Oldenburg, 26111 Oldenburg, Germany
alfred.mertins@uni-oldenburg.de

Abstract. This paper proposes an object-based, highly scalable, lossy-to-lossless coding approach for magnetic resonance (MR) images. The proposed approach, called OBHS-SPIHT, is based on the well known set partitioning in hierarchical trees (SPIHT) algorithm and supports both quality and resolution scalability. It progressively encodes each slice of the MR data set separately in a multiresolution fashion from low resolution to full resolution and in each resolution from low quality to lossless quality. To achieve more compression efficiency, the algorithm only encodes the main object of interest in the input data set, and ignores the unnecessary background. The experimental results show the efficiency of the proposed algorithm for multiresolution lossy-to-lossless MRI data coding. OBHS-SPIHT, is a very attractive coding approach for medical image information archiving and transmission applications especially over heterogeneous networks.

1 Introduction

From the coding point of view, the main features required for an efficient clinical picture archiving and communications systems (CPACS) can be highlighted as follows: efficient lossy-to-lossless compression, object-based functionality and high degree of scalability support.

Volumetric medical images (e.g. MR and CT) are 3D data sets which consist of a sequence of 2D data slices. For efficient archiving and transmission of such vast amounts of data a high degree of compression is required. For instance, an uncompressed typical gray scale MR set of 58 slices of 512×512 resolution results in a data volume of 116 Mbits, and downloading such information via a 56 kbps Internet connection for a remote diagnosis purpose will take more than 35 minutes. For medical image coding lossy-to-lossless compression is required to enable the provision of appropriate services for different applications according to their sensitivity to the image quality in the diagnosis process. Since lossless compression does not degrade the image, it facilitates more accurate diagnosis, of course at the expense of lower compression ratios (i.e. higher bit rates). However, lossy compression is required to significantly reduce transmission and storage costs where the loss is not diagnostically significant.

Over the past decade, wavelet-based image compression schemes have become increasingly important and gained widespread acceptance. An example is the new JPEG2000 still image compression standard [1, 2]. Due to the multiresolution signal representation offered by the wavelet transform, wavelet based coding schemes have a great potential to support scalability features. Among the state-of-the-art embedded wavelet coding approaches, the Set Partitioning in Hierarchical Trees (SPIHT) algorithm [3] is well known as a benchmark for its compression efficiency, full SNR scalability support and very low complexity. These features have made SPIHT very attractive for medical image coding as well [4, 5, 6]. As shown in [4], an object-based version of SPIHT (OB-SPIHT) exhibits a very competitive PSNR performance for the compression of medical images. On the other hand, research conducted by Pearlman [7] showed a very significant complexity reduction of SPIHT over JPEG2000. Although the SPIHT bitstream is tailored for full SNR scalability and is progressive (by quality) coding, which can support lossy to lossless decoding, it does not support spatial scalability to provide a bitstream that can be parsed for multiresolution decoding by different clients with different capabilities.

Often there are regions inside a medical image that contain the main information required for diagnostic purposes. An object-based coding is desirable to enable coding of any region of interest with arbitrary shape in the image, separately from the other parts of the image. This feature helps to achieve a very high compression ratio by only focusing on the important regions in the image and discarding the non-important background that usually takes a large area of medical images, or by encoding the background at a lower precision with a lossy image coder [4, 8]. The region of interest (ROI) coding feature in the JPEG-2000 standard considers the whole image for coding but it applies a higher coding precision to the ROI [9, 10, 11]. On the other hand, an object-based coding makes it possible to encode the ROI as a separate object regardless of the rest of the image.

This research proposes an object-based medical image coding system based on the highly scalable set partitioning in hierarchical trees (HS-SPIHT) algorithm. The HS-SPIHT, introduced by the authors of this paper in their previous works [12, 13], is a modification of the SPIHT algorithm [3] that adds spatial scalability features to the SPIHT algorithm without sacrificing the interesting features of the original algorithm. The coding system proposed in this paper, called OBHS-SPIHT, extends the 2D HS-SPIHT algorithm to object-based coding of MRI data. The OBHS-SPIHT algorithm fulfills all the highlighted requirements for medical image information archiving and transmission systems mentioned earlier in this section.

The rest of this paper is organized as follow. Section 2 gives an overview of the OBHS-SPIHT coding system. In Section 3, the OBHS-SPIHT coding algorithm is presented. The scalable structure of the OBHS-SPIHT bitstream is explained in Section 4. In Section 5, some details about the simulation of the coding system are given and experimental results for multiresolution lossless as well as lossy decoding are presented, and finally, Section 6 concludes the paper.

2 System Overview

The proposed OBHS-SPIHT coding system is depicted in Figure 1. The system input is volumetric MR data set which consists of various slices. On the encoder side, each

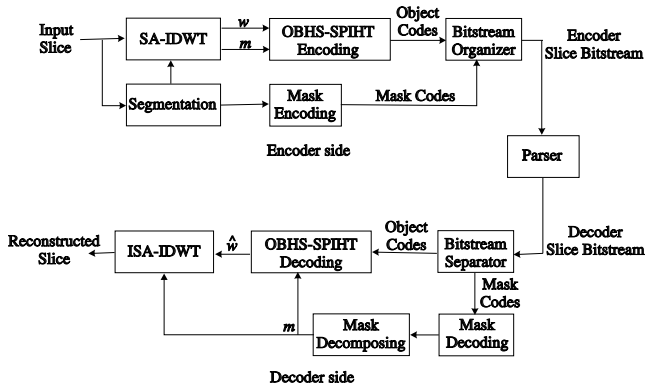


Fig. 1. Block diagram of the OBHS-SPIHT coding system. w denotes the wavelet coefficients, and m means the decomposed mask.

slice is first segmented to extract the medical object of interest from the background. Each voxel in the data set is considered either inside or outside the object. The extracted object is decomposed by a shape-adaptive integer DWT (SA-IDWT) approach which maps integer object voxels to integer wavelet coefficients. Details on the segmentation process and the DWT will be given in Section 5.

The decomposed object coefficients and the decomposed shape mask are then consigned to the OBHS-SPIHT encoder. The encoder only encodes the coefficients that belong to the decomposed object. To recognize these coefficients it uses the decomposed shape mask. The bitstreams from the shape coding and object coding algorithms are assembled in the bitstream organizer to generate the final encoder output bitstream.

In a customization stage, the encoded bitstream is reordered and truncated by a parser which provides proper bitstreams for multiscale lossy-to-lossless decoding. On the decoder side, the bitstream separator first extracts the mask and the object bitstreams from the parsed bitstream. The shape mask is then reconstructed by decoding the shape bitstream. The decomposed mask, which is required by the OHS-SPIHT decoder, is provided by applying the same level of decomposition as used by the encoder to the shape mask. The OHS-SPIHT decoder then decodes the object bitstream, and the inverse SA-DWT is applied to the decoded wavelet coefficients to reconstruct the original slice object at the requested resolution and rate.

3 Object-Based HS-SPIHT

The SPIHT algorithm of [3] considers sets of coefficients that are related through the parent-offspring dependency depicted in Figure 2. In its bitplane coding process, the algorithm deals with the wavelet coefficients as either members of insignificant sets, individual insignificant pixels, or significant pixels. It sorts these coefficients in three ordered lists: the list of insignificant sets (LIS), the list of insignificant pixels (LIP),

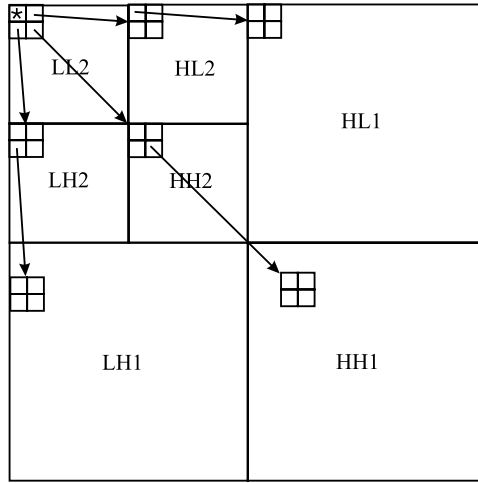


Fig. 2. 2D SPIHT Parent-offspring dependency across wavelet subbands in each slice

and the list of significant pixels (LSP). The main concept of the algorithm is managing these lists in order to efficiently extract insignificant sets in a hierarchical structure and identify significant coefficients, which is the core of its high compression performance. The SPIHT algorithm provides a progressive (by quality) bitstream which is fully SNR scalable, however its bitstream does not support spatial scalability.

In [12, 13] we proposed a scalable modification of SPIHT for image coding, called highly scalable SPIHT (HS-SPIHT), through the introduction of multiple resolution-dependent lists and a resolution-dependent sorting pass. In general, a wavelet decomposed slice with N levels of 2D decomposition enables a scalable encoder to provide at most $N + 1$ different spatial resolution levels. To distinguish between different resolution levels, we denote the lowest spatial resolution level as level $N + 1$. The spatial resolution related to level k is $1/2^{k-1}$ of the resolution of the original data set. The full resolution (the original sequence) then becomes level 1. The three subbands (HL_k, LH_k, HH_k) that need to be added to increase the spatial resolution from Level $k + 1$ to Level k are grouped and called spatial subband set level k . The HS-SPIHT algorithm encodes the different resolution subbands in the wavelet decomposed image separately, allowing a parser or a decoder to directly access the data needed for reconstruction of a desired spatial resolution and/or quality. To manage the scalable coding process, for each resolution subband set, the algorithm defines a set of LIP, LSP and LIS lists, therefore there are $LIP_k, LSP_k,$ and LIS_k for $k = s_{max}, s_{max} - 1, \dots, 1$ where s_{max} is the maximum number of spatial resolution levels supported by the encoder. To improve the algorithm to be used for coding of medical images which contain objects with any arbitrary shape, we only consider and process those coefficients that belong to the decomposed object (see Figure 3) and those sets that are at least partially located inside the decomposed object, similar to the SA-SPIHT algorithms in [14].

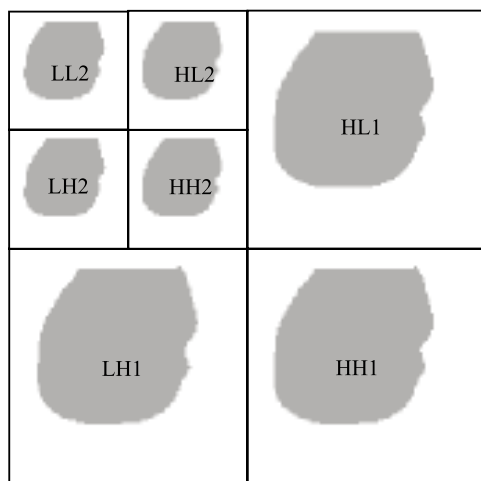


Fig. 3. Example of a decomposed mask of an arbitrarily shaped object

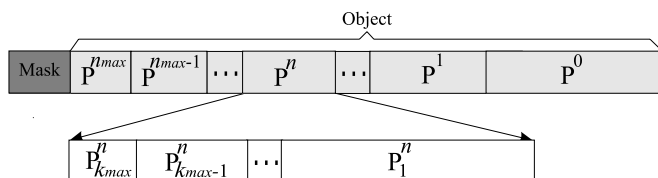


Fig. 4. Structure of the OBHS-SPIHT encoder bitstream for a slice. P_k^n is related to the codepart of spatial subband set level k at bitplane level n .

4 Bitstream Structure

Figure 4 shows the structure of the bitstream generated by the OBHS-SPIHT encoder for a slice. The scalable object bitstream is constructed of different codeparts (P^n), where each part belongs to a bitplane level. Inside each bitplane codepart, the bits that belong to the different spatial subband sets, P_k^n , are separable. To support bitstream parsing, some markers are put in the bitstream to provide the information required for identifying the different resolution and bitplane codeparts in the parsing process.

The encoder needs to encode the input object only once at a lossless rate (covering all biplane coding levels from the maximum bitplane level to bitplane level 0). Different bitstreams for different spatial resolutions can be easily generated from the encoded bitstream by selecting the related resolution codeparts. The parsing process is a simple codeparts-selection procedure and can be carried out by a server that stores the encoded medical data sets or by an individual parser as a part of an active network. The parser does not need to decode any part of the bitstream. As a distinct feature, the reordered bitstreams for each spatial resolution are completely rate-embedded (fine granular at bit

Table 1. Description of the MR data sets used as test volumetric medical images in this paper

History	Age	sex	File name	Voxel size (mm)	Volume size
Congenital heart disease	1	M	MR_ped_chest	$0.78 \times 0.78 \times 5$	$256 \times 256 \times 77$
Normal	38	F	MR_liver_t	$1.45 \times 1.45 \times 5$	$256 \times 256 \times 58$
Normal	38	F	MR_liver_t2e1	$1.37 \times 1.37 \times 5$	$256 \times 256 \times 58$
Left exophthalmos	42	M	MR_sag_head	$0.98 \times 0.98 \times 3$	$256 \times 256 \times 58$

level) and can be truncated at any point up to the level of a perfect lossless reconstruction. Note that the markers in the main bitstream are only used by the parser and do not need to be sent to the decoder.

5 Experimental Results

5.1 Simulation Details

The OBHS-SPIHT coding system were fully software implemented. As volumetric medical data we have chosen the four gray-scale (8 bits per voxel) MR data sets that were also used in [6,5,15]. A description of these MR sets is given in Table 1. To extract the objects from the unimportant, very low magnitude background voxels, a two-stage threshold-based segmentation scheme was used. In the first stage, each MR set was compared with a threshold and all voxels that exceeded the threshold were considered to belong to the object. In a second stage, all background areas that were surrounded by the object were reclassified to belong to the object. The first slice of one of the MR test set, MR_sag_head, and its appropriate segmentation mask is shown in Figure 5. For the object-based wavelet decomposition, an efficient, non-expansive SA-DWT approach, based on the method introduced in [16] was implemented. The integer I(2,2) wavelet filter bank [17] was implemented in a lifting scheme and used for object decompositions with symmetric extension at the boundaries of the object in each slice.

The OBHS-SPIHT encoder was set to progressively encode the decomposed objects of all slices of each MR test set to the lossless rate with three levels of spatial scalability support. The binary mask information for each slice was encoded by an arithmetic binary coding scheme [18].

5.2 Results

Table 2 provides the average bits per voxel (bpv) obtained by OBHS-SPIHT for multiresolution lossless coding of the four MR object sets. As the results show for both cases, a lossless version of the lower resolutions can be obtained at very small rates. Figure 6 shows the lossless reconstruction of slice 9 of MR_sag_head data set at three different resolutions (full, half and quarter). The average rate consumed for coding of the binary mask information of the MR sets lies between 0.016 bpv to 0.02 bpv and therefore negligible.

In Table 3, the OBHS-SPIHT results for lossless coding at full resolution are compared with HS-SPIHT, SPIHT, JPEG2000, JPEG-LS and WinZip coding approaches.

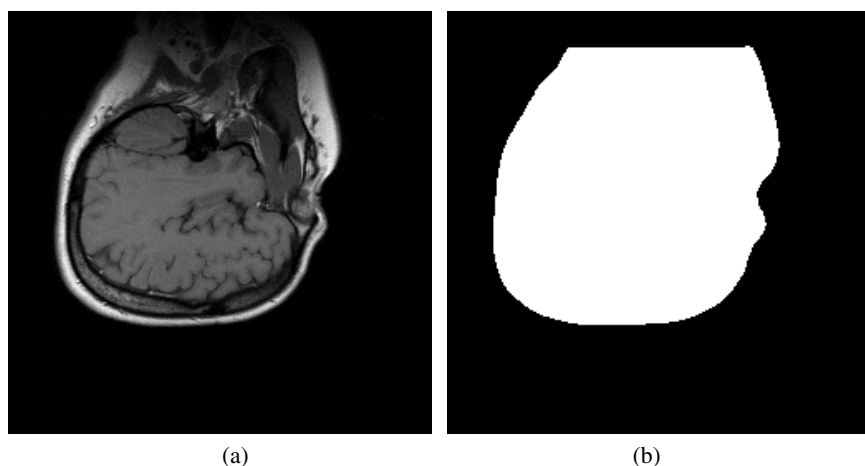


Fig. 5. The first slice of the MR_sag_head data set. (a) Original data. (b) Extracted mask.

Table 2. Average bits per voxel obtained for lossless encoding of the MR data sets by OBHS-SPIHT

Spatial resolution	lossless bits per voxel (bpv)			
	MR_ped_chest	MR_liver_t1	MR_liver_t2e1	MR_sag_head
Quarter	0.1419	0.2722	0.2605	0.1727
Half	0.4339	0.8320	0.8378	0.5435
Full	1.2550	2.3420	2.4955	1.7440

For these coding approaches, the object background in all slices was set to zero to have a fair comparison with OBHS-SPIHT. A very small difference between the lossless compression rates of HS-SPIHT and SPIHT is due to the extra budget consumed by HS-SPIHT for markers in the bitstream which are required for the parsing process. The results reported here for SPIHT, HS-SPIHT and OBHS-SPIHT were obtained without extra arithmetic coding of the encoder output bitstreams. As shown in [3], an improved coding performance for SPIHT and consequently for HS-SPIHT can be achieved by further compressing the binary bitstreams with an arithmetic coder. Despite this fact, the OBHS-SPIHT algorithm provides comparable results to JPEG2000 while it has much less complexity [7]. As the results show, JPEG-LS outperforms the other coders, but it does not support spatial scalability and its bitstream can not be used for lossy decoding.

To show the full scalability of OBHS-SPIHT, Table 4 presents some numerical results for multiresolution decoding of the MR test sets at a wide range of bit rates. This is based on a scenario of one-time-encoding and multiple-times-decoding, by parsing the encoder bitstream for various resolutions and rates, which is required for serving different clients with different capabilities in archiving and transmission systems especially over a heterogeneous system like the Internet. In such systems each client can request a specific bit rate and resolution level which fits its needs.

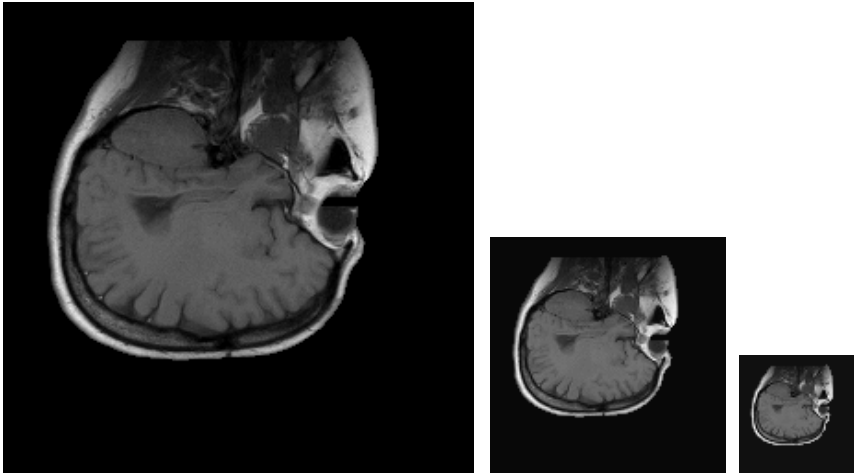


Fig. 6. Lossless reconstruction of slice 9 of MR_sag_head at full, half and quarter resolution by OBHS-SPIHT decoder

Table 3. Comparison of average bits per voxel obtained for lossless encoding of the MR data sets at full resolution with different coding methods

Method	MR_ped_chest	MR_liver_t1	MR_liver_t2e1	MR_sag_head
OBHS-SPIHT	1.2550	2.3420	2.4955	1.7440
HS-SPIHT [12]	1.5921	2.6354	2.7781	2.1772
SPIHT [3]	1.5818	2.6247	2.7677	2.1660
JPEG2000 [19]	1.4537	2.2266	2.3499	1.9029
JPEG-LS [20]	1.2183	1.9587	2.1134	1.5911
WinZip	1.8900	3.7261	3.7512	2.3571

Table 4. PSNR results for lossy decoding of the OBHS-SPIHT bitstreams at different spatial resolutions and rates

Spatial resolution	rate (bpv)	PSNR (dB)			
		MR_ped_chest	MR_liver_t1	MR_liver_t2e1	MR_sag_head
Quarter	0.0625	45.72	35.65	35.59	42.84
	0.125	58.26	45.37	45.58	53.70
Half	0.0625	33.56	28.79	27.62	32.70
	0.125	40.43	33.88	31.96	38.45
	0.25	48.00	40.26	38.85	44.75
Full	0.125	32.42	28.13	25.67	30.55
	0.25	36.72	32.96	29.90	34.33
	0.5	42.23	37.03	34.89	38.43
	1	47.70	43.05	40.24	43.47

6 Conclusions

An object-based, highly scalable wavelet coding system, OBHS-SPIHT, for lossy-to-lossless coding of MR data was presented. The object of interest in each slice of MR data sets were segmented from the background. A reversible shape-adaptive integer DWT was used to decompose the input objects. Each slice of the data set was encoded separately. This not only facilitates more efficient random access to the slices, but also requires less memory from the coding system. The OBHS-SPIHT bitstream is easily reorderable by a simple parser for multiresolution decoding. The experimental results for lossy and lossless cases on some MR data sets at various spatial resolution levels showed the excellent performance of the proposed algorithm. Possessing important features such as arbitrarily shaped object coding and full resolution and quality scalability functionalities makes the proposed approach attractive for volumetric medical image information archiving and transmission systems.

References

1. Taubman, D.S., Marcellin, M.W.: *JPEG2000 : Image Compression Fundamentals, Standards, and Practice*. Kluwer, Boston, MA (2002)
2. Christopoulos, C.: *JPEG-2000 verification model 8.5 (technical description)*. Technical report (2000) *ISO/IEC JTC1/SC 29/WG1 N1878*.
3. Said, A., Pearlman, W.A.: A new, fast and efficient image codec based on set partitioning in hierarchical trees. *IEEE Trans. Circ. and Syst. for Video Technology* **6** (1996) 243–250
4. Penedo, M., Pearlman, W., Tahoces, P.G., Souto, M., Vidal, J.J.: Region-based wavelet coding methods for digital mammography. *IEEE Trans. Medical Imaging* **22** (2003) 1288–1296
5. Kim, Y., Pearlman, W.A.: lossless volumetric medical image compression. In: *Proc. SPIE*. Volume 3808. (1999) 305–312
6. Cho, S., Kim, D., Pearlman, W.A.: Lossless compression of volumetric medical images with improved 3-D SPIHT algorithm. *Journal of Digital Imaging* **17** (2004) 57–63
7. Pearlman, W.A.: Trends of tree-based, set-partitioning compression techniques in still and moving image systems. In: *Proc. Picture Coding Symposium (PCS'2001)*, Seoul, Korea (2001) 1–8
8. Menegaz, G., Thiran, J.P.: Lossy to lossless object-based coding of 3-D MRI data. *IEEE Trans. Image Processing* **11** (2002) 1053–1061
9. Anastassopoulos, G.K., Skodras, A.N.: JPEG2000 ROI coding in medical imaging applications. In: *Proc. Second IASTED Int. Conf. Visualization, Imaging and Image Processing*, Anahim, CA, USA, ACTA Press (2002) 783–788
10. Christopoulos, C., Askelof, J., Larsson, M.: Efficient methods for coding regions of interest in the upcoming JPEG2000 still image coding standard. *IEEE Signal Processing Lett.* **7** (2000) 247–249
11. Christopoulos, C., Askelof, J., Larsson, M.: Efficient regions of interest encoding techniques in the upcoming JPEG2000 still image compression standard. In: *Proc. Int. Conf. Image Processing (ICIP)*. Volume 2., Vancouver, BC, Canada (2000) 41–44
12. Danyali, H., Mertins, A.: Highly scalable image compression based on SPIHT for network applications. In: *Proc. IEEE Int. Conf. Image Processing (ICIP'2002)*. Volume 1., Rochester, NY, USA (2002) 217–220
13. Danyali, H., Mertins, A.: Fully spatial and SNR scalable, SPIHT-based image coding for transmission over heterogenous networks. *Journal of Telecommunications and Information Technology* (2003) 92–98

14. Minami, G., Xiong, Z., Wang, A., Mehrotra, S.: 3-D wavelet coding of video with arbitrary regions of support. *IEEE Trans. Circ. and Syst. for Video Technology* **11** (2001) 1063–1068
15. Bilgin, A., Sementilli, P.J., Sheng, F., Marcellin, M.W.: Scalable image coding using reversible integer wavelet transforms. *IEEE Trans. Image Processing* **9** (2000) 1972–1977
16. Li, S., Li, W.: Shape-adaptive discrete wavelet transforms for arbitrarily shaped visual object coding. *IEEE Trans. Circ. and Syst. for Video Technology* **10** (2000) 725–743
17. Calderbank, A., Daubechies, I., Sweldens, W., Yeo, B.L.: Wavelet transforms that map integers to integers. *Appl. Comput. Harmon. Anal.* **5** (1998) 332–369
18. Brady, N., Bossen, F., Murphy, N.: Context-based arithmetic encoding of 2D shape sequences. In: *Proc. IEEE Int. Conf. Image Processing (ICIP' 1997)*. Volume 1., Santa Barbara, CA, USA (1997) 29–32
19. JJ2000 (An implementation of JPEG2000 standard in Java) Version 4.1 available at: http://jj2000.epfl.ch/jj_download/index.html.
20. Weinberger, M.J., Seroussi, G., Sapiro, G.: LOCO-I: a low complexity, context-based lossless image compression algorithm. In: *Proc. IEEE Data Compression Conference*, New York (1996) 140–149

A Novel Fuzzy Segmentation Approach for Brain MRI

Gang Yu¹, Changguo Wang², Hongmei Zhang¹, Yuxiang Yang¹,
and Zhengzhong Bian¹

¹ School of Life Science and Technology, Xi'an Jiaotong University,
Xi'an 710049 China

yugang@mailst.xjtu.edu.cn

² Nantong Vocational College, Nantong 226007 China

Abstract. A novel multiresolution approach is presented to segment Brain MRI images using fuzzy clustering. This approach is based on the fact that the image segmentation results should be optimized simultaneously in different scales. A new fuzzy inter-scale constraint based on antistrophic diffusion linkage model is introduced, which builds an efficient linkage relationship between the high resolution images and low resolution ones. Meanwhile, this paper develops two new fuzzy distances and then embeds them into the fuzzy clustering algorithm. The distances describe the fuzzy similarity in adjacent scales effectively. Moreover, a new multiresolution framework combining the inter- and intra-scale constraints is presented. The proposed framework is robust to noise images and low contrast ones, such as medical images. Segmentation of a number of images is illustrated. The experiments show that the proposed approach can extract the objects accurately.

1 Introduction

Multiscale or multiresolution approaches for medical image analysis, such as the pyramid[1], stack[2] and wavelets[3], have gained considerable attention. The segmentation in conventional pyramid[1] is accomplished by a downward projection from one scale to another scale, which limits the possible number of segments to $4i (i \in N)$. The stack is a successful multiresolution method in 2D images, but the transition from 2D to 3D is far from trivial. Keon et al developed a new multiscale image segmentation technique[4-6], i.e. the hyperstack. The conventional (single-parent) hyperstack is characterized by the fact that a voxel at one level of the hyperstack is connected to at most one (parent) voxel in next higher layer. The extension, probabilistic (multiparent) hyperstacks, is introduced [6], in which children are allowed to link to multiple parents, but its computational cost is very expensive. In the hyperstack segmentation method, many linkage criteria are proposed to build the most possible child-parent relationship, which demonstrates the similarity between the voxels at adjacent level efficiently.

Some fuzzy clustering approaches based on multiresolution framework were proposed recently[7][11-15]. Mahmoud et al proposed an efficient method [7],

but the fuzzy clustering was only regarded as a post-processing step for the over-segmented regions. Punya presented a multiresolution fuzzy clustering algorithm[11], but the Fuzzy C-Means algorithm (FCM) was only implemented on each scale respectively, and neglected the relationship between different scales. A multiresolution color image segmentation approach was presented [12], where a multiscale dissimilarity measure was proposed to measure the inter-region relations. Reference [13] presented a new unsupervised multiresolution pyramidal edge detector, and a multiresolution clustering method was developed to speed up the conventional fuzzy algorithm[14]. Reference [15] described a multiresolution-based approach combined with wavelet analysis. A local FCM segmentation generated an estimate of local intensity. These approaches often apply the fuzzy clustering algorithm directly in each level of scale images respectively, but they fail to introduce the inter-scale relationship for optimizing the intra-scale segmentation. Therefore, the approaches are not enough robust to the degraded images.

This paper develops a novel fuzzy segmentation approach based on anti-strophic diffusion linking model. The first step of the approach is to build the relationship of child-parent scales. The fuzzy clustering combining the inter-scale and intra-scale constraints is then applied for image segmentation. The paper is organized in the following way. In section 2, the multiresolution linking model is described. In section 3, the inter-scale constraint and the multiresolution algorithm are presented. In section 4, experimental results are provided, and conclusions are reported in the end.

2 Nonlinear Diffusion Linking Model

In this section, we describe the nonlinear diffusion linking model, which is constructed as the scale space. The nonlinear diffusion linking model is similar to hyperstack, but root labeling and downward projection applied in hyperstack segmentation are excluded, we only use the proposed multiresolution fuzzy clustering to segment images.

2.1 Blurring and Subsampling

Perona and Malik showed that a scale-space could be represented by a progression of images computed by the heat diffusion equation[8][9][10]. When diffusion coefficient D is defined as a constant in all locations, the diffusion equation is equal to isotropic diffusion, i.e. Gaussian blurring. When D is a matrix, the equation is anisotropic diffusion. The pixel values at high level may be computed by successively applying diffusion equation and then subsampling. Perona and Malik firstly introduced non-linear diffusion, i.e. PM equation [9], within the image processing context. The original image L_0 is blurred by the diffusion equation with a specific diffusion coefficient D suitable for the image pattern, and a coarse image is obtained. The coarse image is then subsampled, and the higher-level image L_1 is obtained. Similarly, more levels can be obtained.

2.2 Linking

In the linking step, the parent-child relationship between any two adjacent layers is defined. Meanwhile, the spatial relationship, between the image elements of two successive layers of the scale space, is always known. For example: in a 2×2 subsample scale space, a pixel in higher level is the parent of four nearest image elements (children) in lower level, so each child has only one parent and therefore no ambiguity exists in the spatial parent-child relationship. Here, this parent is called explicit parent, because it is exclusive. Similarly, the four children are called explicit children.

However, this relationship is ambiguous or fuzzy in a linked model such as hyperstack, where the children of a level can belong to different parents in upper level. The similarity between a child image element and its possible parents is defined to describe how similar they are. Two usual similarity criteria for linking were presented [6]. The first similarity term is based on intensity proximity between children and parents. The second similarity term encourages the convergence to ever fewer parents. The parents are selected on the basis of their affection to a given child. The potential parent with the highest affection value is selected to be the child's parent. This affection is defined as:

$$L(x, y) = \sum_{i=1}^2 \omega_i S_i(x, y),$$

where x and y are a given child and potential parent respectively, ω_i is weight values, $S_i(x, y)$ is the two similarity term.

3 Multiresolution Fuzzy Clustering

3.1 Self-similarity and Constraints of Inter- and Intra-scale

There is self-similarity in a series of images of scale space, because all of them are the approximate representation of original image with different scales.

(1) The similarity in a scale. The similarity in a scale shows obvious clustering features, where the children belonging to identical class are similar. The conventional FCM is competent to describe the similarity, but the similarity in a scale is very sensitive to noise.

(2) The similarity between two successive scales. The children are related to their parents, and the children both inherit the features of their parents, and show some new features. The inherited features include intensity, gradient, and fuzzy clustering.

In order to better describe the relations, some mathematical symbols are introduced. Let $X^{(L)} = \{X_k^{(L)} | k \in I^{(L)}\}$ be the image (or feature image of the image) in level L . $x_k^{(L)}$ is the image value or feature vector of the pixel k , $I^{(L)}$ is the data set. The labeled image is denoted by $l^{(L)} = \{l_k^{(L)} | k \in I^{(L)}\}$, where $l_k^{(L)} \in \{1, 2, \dots, c\}$ represents the label of the pixel k , c is the number of

clustering. The multiresolution segmentation is described as: given $X^{(L+1)}$ and $l^{(L+1)}$ in the higher level $L + 1$, the optimal estimation about $l^{(L)}$ should obey the self-similarity of both inter- and intra-scale. Let $P(x_k^{(L)})$ be defined as the parent of the pixel k in level L according to the linking model, and $PS(x_k^{(L)})$ be the explicit parent of pixel k according to the spatial relationship. $P(x_k^{(L)})$ is defined as:

$$P(x_k^{(L)}) = \operatorname{argmax}\{L(x_m^{(L+1)}, x_k^{(L)})\} \quad m \in N_p(PS(x_k^{(L)})) \quad (1)$$

Where $N_p(PS(x_k^{(L)}))$ is defined as the neighbors of $PS(x_k^{(L)})$, $L(x, y)$ is the affection value between x and y described in the above section. The equation (1) is straightforward. The linked parent $P(x_k^{(L)})$ should be the pixel with maximum affection value in the neighbors of the explicit parent obtained by the spatial relationship. The neighbors decide the search volume of potential parents. For example, the 4-neighbors or 8-neighbors is usual search volume. Similarly, the linked child can be obtained from the linking relationship. Let $S(x_k^{(L+1)})$ be the most possible child of pixel $x_k^{(L+1)}$:

$$S(x_k^{(L+1)}) = \operatorname{argmax}\{L(x_k^{(L+1)}, x_m^{(L)})\} \quad m \in N_c(x_k^{(L+1)}) \quad (2)$$

Where $N_c(x_k^{(L+1)})$ denotes the explicit children of pixel $x_k^{(L+1)}$. For example, every pixel in the level $L + 1$ has four explicit children in the level L , if the 2×2 subsample is applied in the construction of the linking model. $S(x_k^{(L+1)})$ is the child with the maximum affection value in the four children.

According to the self-similarity described above, the fuzzy distances should include two parts: (1) the fuzzy distance in a scale. (2) the fuzzy distance between two adjacent scales. The intra-scale distance is defined in the conventional FCM. The similarity between adjacent scales shows that the fuzzy clustering, the parent and its children belong to, is similar. Moreover, the clustering centers in two adjacent levels are also close. Therefore, two inter-scale fuzzy distances, i.e. $\|P(x_k^{(L)}) - v_i^{(L+1)}\|$ and $\|v_i^{(L)} - S(v_i^{(L+1)})\|$, are introduced, where $v_i^{(L)}$ is the clustering center in Level L . In figure 1, the dashed line represents the parent-child relationship based on the linking model. The first fuzzy distance shows the distance between the linked parent $P(x_k^{(L)})$ of a pixel $x_k^{(L)}$ and the corresponding clustering center $v_i^{(L+1)}$ in the higher level. The second fuzzy distance describes the distance between two corresponding clustering centers of the adjacent levels.

Firstly, the two distances are based on the fact that the parent and child should belong to the two clustering centers with the parent-child relationship respectively, which shows the similarity of the parent-child pixels in different scales. Secondly, two distances also show the fact that the center $v_i^{(L)}$ in low scale L should not be far from the child of the corresponding center $v_i^{(L+1)}$ in the higher scale, which show the similarity of clustering centers in different scales. It is obvious that these facts are similar to the observed result in successive scale

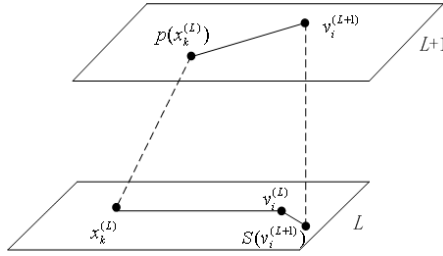


Fig. 1. The inter-scale fuzzy distance

by human eyes. The two distances should be minimized while the fuzzy clustering converges in a global optimal solution. The inter-scale constraint between the current level L and higher Level $L + 1$ is defined as follows:

$$J_m^{(L,L+1)}(U; V; X) = \sum_{k=1}^n \sum_{i=1}^c u_{ik}^{(L)m} [\alpha \cdot \|P(x_k^{(L)}) - v_i^{(L+1)}\|^2 + \beta \cdot \|v_i^{(L)} - S(v_i^{(L+1)})\|^2] \tag{3}$$

Where α, β are parameters, which control the sensitivity of inter-scale constraint. The membership value u_{ik} defines the grade of a feature point x_k belonging to the cluster center v_i . m is a parameter ranging from 1 to ∞ , which controls the fuzziness of the resulting. In most cases, m is set to be 1.5. L and $L + 1$ denote the adjacent levels in the scale space. n and c are the number of feature points and fuzzy clustering in the image respectively. According to FCM, the intra-scale constraint in Level L is defined as follows:

$$J_m^{(L)}(U; V; X) = \sum_{k=1}^n \sum_{i=1}^c u_{ik}^{(L)m} \cdot [\omega \|x_k^{(L)} - v_i^{(L)}\|^2] \tag{4}$$

Where ω is a parameter. Integrate (3) and (4), the multiresolution energy function combining the inter- and intra-constraint is defined as:

$$E(U^{(L)}, V^{(L)} | U^{(L+1)}, V^{(L+1)}) = J_m^{(L)} + J_m^{(L,L+1)} \tag{5}$$

3.2 Multiresolution Fuzzy Segmentation Algorithm

The segmentation begins from the top level, where the pre-segmentation is performed by a conventional clustering method, such as FCM. The result is used to the segmentation in the lower level. The optimal problem in every level is to minimize the energy function described above:

$$(U^{(L)}, V^{(L)}) = \operatorname{argmin} E, \quad \sum_{i=1}^c u_{ik}^{(L)} = 1 \quad \text{for all } x_k \tag{6}$$

From Lagrange method, let $J = E - \lambda \cdot (\sum_{i=1}^c u_{ik}^{(L)} - 1)$. The optimal solution is obtained from the equations: $\partial J / \partial u_{ik}^{(L)} = 0$ and $\partial J / \partial v_i^{(L)} = 0$:

$$u_{ik}^{(L)} = \frac{(\frac{1}{H})^{\frac{1}{m-1}}}{\sum_{i=1}^c (\frac{1}{H})^{\frac{1}{m-1}}} \tag{7}$$

where $H = [\omega \|x_k^{(L)} - v_i^{(L)}\|^2 + \alpha \cdot \|P(x_k^{(L)}) - v_i^{(L+1)}\|^2 + \beta \cdot \|v_i^{(L)} - S(v_i^{(L+1)})\|^2]$

$$v_i^{(L)} = \frac{\sum_{k=1}^n [u_{ik}^{(L)m} \cdot (\omega \cdot x_k^{(L)} + \beta \cdot S(v_i^{(L+1)}))]}{\sum_{k=1}^n [u_{ik}^{(L)m} \cdot (\beta + \omega)]} \tag{8}$$

The equation (7) and (8) are the iteration equations of multiresolution fuzzy clustering in level L . After several iterations, $u_{ik}^{(L)}$ and $v_i^{(L)}$ are the optimal estimation of membership value and fuzzy clustering center in level L respectively.

4 Experiments

We designed two groups of experiments in this section, one for synthetic experiment and the other for MRI brain images. We build the linking model by the conventional PM diffusion equation before the segmentation. In the first group, a 123*112 synthetic image is drawn, where a target and background represent conventional 2-class segmentation. The gray value of target is 186 and the background is 125, and 30% Gaussian noise is added to the image.

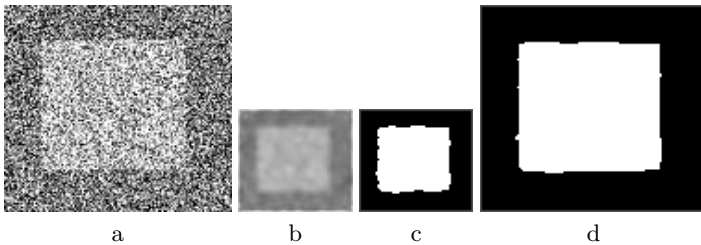


Fig. 2. The segmentation of image with serious noise. (a) is the original image. (b) is the blurred image in high scale, (c) is the segmentation result in the high scale. (d) is the result in the original image by the proposed model.

Figure2 is the experiment result with serious noises. The segmentation result is described by two colors, white and black. From Figure2(d), the target is extracted from the background successfully by our model, only the boundary has a little drawback, because the serious noise spoiled it.

The MRI images are obtained from the McGill Brain Web Database, where the different slices of brain images with different noises are provided. The noise in these datasets varies from 0% to 9%. We downloaded the datasets and segmented

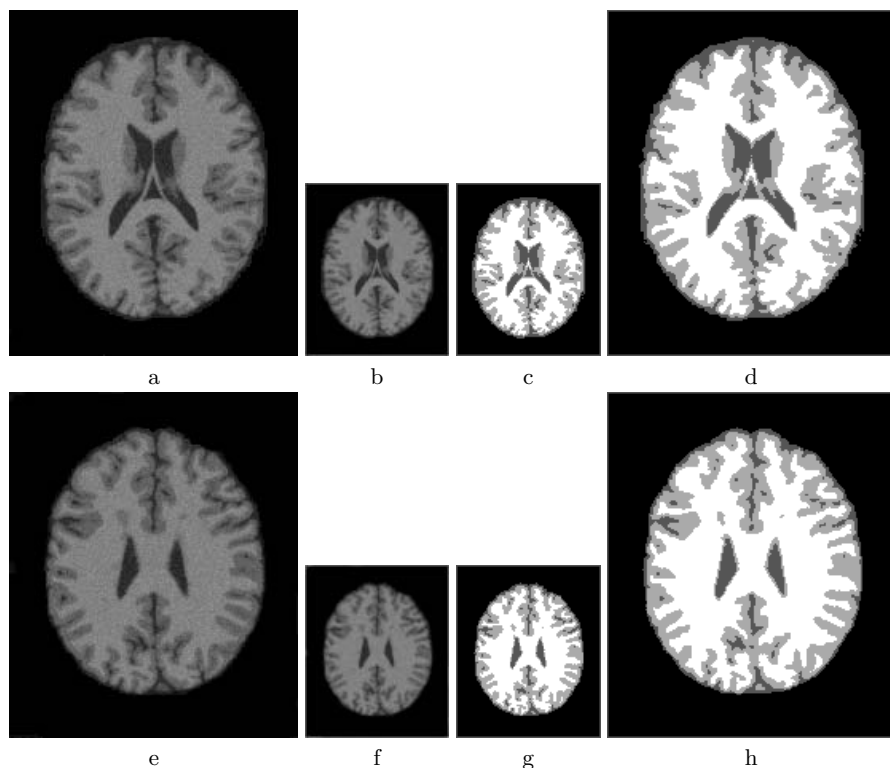


Fig. 3. The segmentation of slice 91 and 100 with 5% noise. (a) slice 91, (e) slice 100. (b), (f) are the coarse images in the high scale. (c), (g) are the segmentation in coarse images. (d) and (h) are the final segmentation results.

the slices for extracting three matters, i.e. brain gray matter, white matter and cerebrospinal fluid.

Figure 3 describes the segmentation of MRI images with 5% noise, where gray matter, white matter and cerebrospinal fluid are extracted successfully. Figure 3(d) and (h) demonstrate the performance of our model in the low contrast medical images.

Figure 4 describes the segmentation of MRI images with 9% noise, which is also the biggest noise in the brain databases. The results of Figure 4 (d) and (h) are similar to those of Figure 3 (d) and (h), but a little gray matter is wrongly classified as white matter, and a little cerebrospinal fluid is assigned to gray matter. Even so, the segmentation has a satisfying result as to these degraded MRI images.

The selection of parameters is described in table 1, where the parameters should be chosen according to the noise degree. Usually, the parameter ω is fixed to 1, while other parameters are proportionally chosen to control the weights between the inter- and intra-scale constraints.

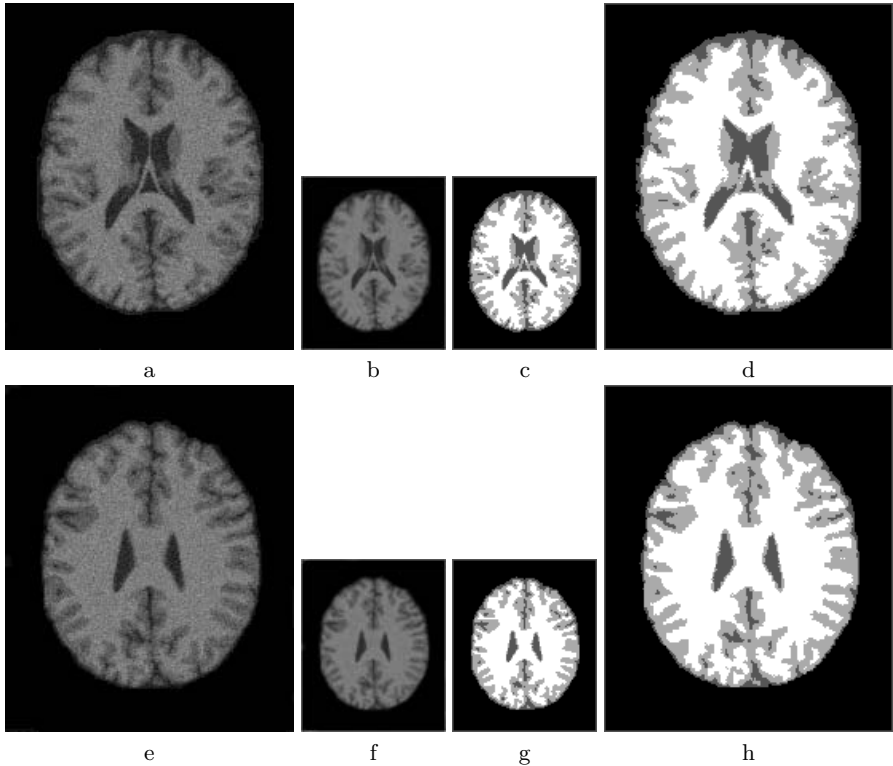


Fig. 4. The segmentation of slice 91 and 100 with 9% noise. (a) slice 91, (e) slice 100. (b) , (f) are the coarse images in the high scale. (c), (g) are the segmentation in coarse images. (d) and (h) are the final segmentation results.

Table 1. The selection of parameters in the experiments

Noise	ω	α	β
5%	1	0.5	1
9%	1	1	1
15%	1	5	2
30%	1	8	2

Where $\alpha \in [0.5, 10]$, $\beta \in [1, 3]$ represent the confidence in the inter-scale constraints. When the noise in the images becomes more serious, α, β should increase accordingly. After many experiments, we find that α, β are not sensitive to the segmentation results because a little change of the parameters hardly influences the final results. In most cases, the value of α, β can be set as an integer, where the increment is 1. The selection of parameters refers to the Table 1.

The computation efficiency of proposed model is very desirable, which is close to conventional FCM, because the re-sampling in the high scale reduces the datasets evidently. Meanwhile, the segmented clustering centers in high scale

can be used as the initial condition of low scale image according to the linkage relationships, which also reduces the iteration times for the final convergence. Every computation time of the above experiments approximates to 1 second, where our model is programmed by Matlab 6.5 in the computer with CPU P4-1.6GHz and 256M memory.

5 Conclusion

This paper proposed a novel multiresolution fuzzy method for MRI image segmentation. This proposed approach is based on the fact that the image segmentation results should be optimized simultaneously in different scales. In this approach, the similarity between adjacent scales was built by a nonlinear linking model. A new inter-scale fuzzy constraint was then introduced. The constraint described an efficient fuzzy linkage relationship between the high scale and low scale. Two fuzzy distances defined based on the constraint showed the similarity of parent-child pixels and clustering centers in successive scales. We developed a new energy function and then embedded it into the conventional fuzzy clustering. Meanwhile, new fuzzy clustering iteration equations were derived, which was utilized for computing the fuzzy partition matrix at different resolutions.

The approach is robust to noise images and low contrast ones, because the optimization is applied in different scale. We segmented several images including synthetic image and Brain MRI images with different noise. Segmentation results showed that the proposed approach is accurate for extracting the objects. Moreover, the approach is not only used to segment MRI images, but also for usual pattern classification. Besides, the computation time is very desirable, which is competent to high performance of real application.

References

1. M. Bister, J. Cornelis, A. Rosenfeld.: A Critical View of Pyramid Segmentation Algorithms Pattern Recognition Letters, Vol.11, (1990)605-617.
2. L.M. Lifshitz, S.M. Pizer.: A Multiresolution Hierarchical Approach to Image Segmentation Based on Intensity Extrema. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 12, No. 6, (1990)529-541.
3. S.G. Mallat.: A Theory for Multiresolution Signal Decomposition: The Wavelet Representation. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 11, No. 7, (1989)674-694.
4. K.L. Vincken.: Probabilistic Multiscale Image Segmentation by the Hyperstack. PhD thesis, Utrecht Univ., The Netherlands(1995).
5. K.L. Vincken, C.N. de Graaf, A.S.E. Koster, M.A. Viergever, F.J.R. Appelman, and G.R. Timmens.: Multiresolution Segmentation of 3D Images by the Hyperstack. Proc. First Conf. Visualization in Biomedical Computing, (1990)115-122.
6. Koen L. V.: Probabilistic Multiscale Image Segmentation. IEEE Trans. Pattern Analysis and Machine Intelligence, Vol. 19, No. 2, (1997)109-120.
7. Mahmoud R. R., Pieter M. J., Boudewijn P. F., et al.: A Multiresolution Image Segmentation Technique Based on Pyramidal Segmentation and fuzzy Clustering. IEEE Transactions on Image Processing, Vol. 9, No. 7, (2000)1239-1248.

8. Weickert J.: Coherence-enhancing diffusion of colour images. *Image and Vision Computing*, Vol 17, (1999)201-212.
9. Perona P, Malik J.: Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Machine Intelligence*, Vol 12, No 7, (1990)629-639.
10. S. T. Acton, A. C. Bovik, M.M Crawford.: Anisotropic Diffusion Pyramids for Image Segmentation. *IEEE Conference on Image Processing*, (1994)478-482.
11. Thitimajshima, P.: Multiresolution Fuzzy Clustering for SAR image Segmentation. *Geoscience and Remote Sensing Symposium. IGARSS '99 Proceedings*, Vol5, (1999)2507 - 2509.
12. Sokratis M.: Segmentation of Color Images Using Multiscale Clustering and Graph Theoretic Region Synthesis. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans*, Vol. 35, No. 2, (2005)224-238.
13. Zhigang W., Dong W., , Wei W., Xiaoming X.: Pyramidal Edge Detector Based on Adaptive Weighted Fuzzy Mean Filters. *Proceedings of SPIE - The International Society for Optical Engineering*, Vol 4550, (2001)46-51.
14. M. Buerki, K.O. Lovblad, H. Oswald et al.: Multiresolution Fuzzy Clustering of Functional MRI Data. *Neuroradiology*, Vol. 45, (2003)691-699.
15. S. Sentelle C. Sentelle and M.A. Sutton.: Multiresolution-Based Segmentation of Calcifications for the Early Detection of Breast Cancer. *Real-Time Imaging* 8, (2002)237-252.

Extrema Temporal Chaining: A New Method for Computing the 2D-Displacement Field of the Heart from Tagged MRI

Jean-Pascal Jacob¹, Corinne Vachier¹, Jean-Michel Morel¹,
Jean-Luc Daire², Jean-Noel Hyacinthe², and Jean-Paul Vallée²

¹CMLA (Centre des Mathématiques et de Leurs Applications) of ENS (Ecole Normale Supérieure) Cachan

² Project of the Swiss National Science Foundation PP00B-68778/1, Department of radiology of the GUH (Geneva University Hospital)

Abstract. This work takes is part of a medical research project which intends to induce and study cardiac hibernation in rats. The underlying goal is to understand the physiology of heart disease. We present here a novel method to compute the 2D-deformation field of the heart (rat or human) from tagged MRI. Previous work is not suitable for wide clinical use for different reasons, including important computing time and lack of robustness. We propose an original description of tags as local minima of 1D signals. This leads us to a new formulation of the tag tracking problem as an Extrema Temporal Chaining (ETC) and a 2D-rendering. 2D-displacements are then interpolated on a dense field. The developed method is fast and robust. Its performances are compared to those of HARP, a leading method in this field.

Introduction

The present work is part of a medical research project that is carried out in the department of radiology of the GUH². It attempts to induce and study the cardiac hibernation phenomenon in rats. For this purpose, the analysis of the heart's inner-wall motion is fundamental.

Among the noninvasive methods available for studying the heart motion, tagged magnetic resonance imaging techniques [3,4] enable the visualization of inner-wall motion : it consists of adding a regular pattern at the beginning of a normal cine-MRI (Magnetic Resonance Imaging) sequence. The CSPAMM (Contrast enhanced SPAtial Modulation of Magnetization [3,5]) tagging technique produces the most usual pattern : 2 orthogonally tagged sequences with a sinusoidal-like profile, yielding alternating white and black stripes (see Fig.1). During the acquisition, the tag pattern deforms with the tissue, allowing the tracking of motion of points within the myocardium. The dark lines are called tags or tag lines. An example of tagged MRI sequence in the case of a human heart is presented in Fig. 1. The same imaging acquisition process was successfully adapted to rat imaging on a Philips 1.5 T scanner, but with a lower quality due to physical limitations.

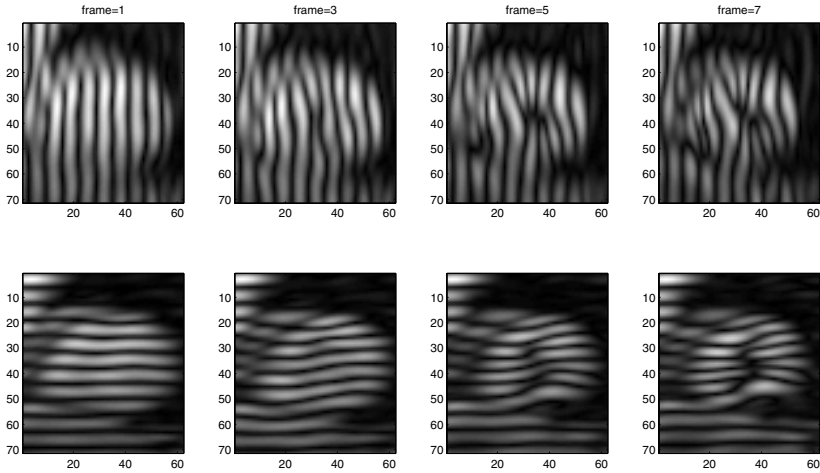


Fig. 1. 4 images of a human heart, from a tagged MRI CSPAMM sequence in both tagging directions, frames 1, 3, 5, 7. The image size is 71×62 pixels, the tag spacing equals 6 pixels, and the inter-image delay is about 36 ms. To overcome the visual effect of the fading in the figure, the contrast has been linearly enhanced.

Beyond the medical qualitative analysis, clinicians wish to measure precisely the displacement of the myocardium points, including points at the border of the cardiac muscle where the signal is noisiest. This is especially important to extract myocardial strains that accurately measure cardiac motion. A manual exploitation of the tagged images is not desirable since it would take many hours per sequence [4].

Several estimations of the myocardium motion have been proposed in the literature during the past twenty years. They are based on ideas such as tag tracking, optical flow computing, or signal phase analysis. Two softwares are most prominent : FindTags [9] developed by M.A.Guttman, J.L. Prince and E.R.McVeigh, and HARP [27] developed by N.Osman.

Why not work with one of the existing softwares ? Because none of the methods developed until now in the tagged MRI field has been extensively validated in clinical context. The HARP method is certainly the most advanced in that sense, but it is not yet operational : in particular, it suffers from coarse tag jump errors and some local instabilities near the myocardium boundaries. Findtags, in turn, is difficult to use and not widely available.

One major difficulty when analyzing tagged MRI is the poor resolution and contrast present in the images. In our context this is accentuated for the rat studies. Thus the goal of our work is to build a robust, accurate, fast and fully automatic procedure of analysis and to validate it in a clinical context.

The paper is organized as follow: predominant works on tagged MRI are recalled in section 1. The extrema temporal chaining method is described in

section 2. Section 3 deals with the 2D deformation field. Section 4 is devoted to a comparison of the results obtained with our ETC method and with the HARP method.

1 Previous Work on Tagged MRI

Previous work on tagged MRI is of two types : solutions based on tag detection and tracking (the oldest ones), and solutions based on derivatives calculation. We describe some of them in the 2 next paragraphs.

Many tag detection methods have been presented in the literature [6,12,9,8,7]. The three most popular ones are the following ones : that of Fisher [6], based on a 2D-correlation of the grid nodes (intersections). A dynamic post-processing is then used to track the tag nodes through time [17]. The FINDTAGS method of Guttman [9,10] allows to find and track tags through time via a template matching of the physical expected tag profile. These two methods require a priori knowledge on the tagging sequence, and lead to errors near the myocardium boundaries. Finally, we mention the snake tracking of the tag lines, first proposed by Young [12], Amini [13], and Kumar [11]. Major drawbacks of the snake-based solutions are that they require a prior initialization of the tag position in the first frame, and that they are often time consuming. These three methods are highly parameterized, especially due to the fading effect which leads to parameter variations.

Tag tracking is not a goal in itself but a way to compute a dense deformation vector field. Most currently, only the tags intersections are used [30] since the velocity may be directly computed on those points with analytical [31] or stochastic multidimensional models [32]. But it is also possible to use the plain tracked shapes information [14].

The second category of methods propose a direct calculation of the deformation vector field. In comparison with the precedent ones, based on the tags detection, the plain information of the image is used (and not only the subset of the image corresponding to the tags). As a consequence, these methods have fewer parameters. These methods are of three types :

1. the optical flow methods, which have to deal with the fading effect in different manners [22,21,25,23,24]. In these methods, only the grid images are used. It means that the computation is precise only at points on the grid, which implies a weak precision of the values. A direct calculation without any model could be derived from 2 CSPAMM sequences.
2. the phase constancy based method, known as HARP (HARmonic Phase images). The HARP method [27] uses the phase invariance of a material point. The phase map is computed via a simple Fourier filtering followed by an unwrapping procedure. HARP uses the two orthogonally tagged sequences [26] for a direct calculation of the velocity field. The results are visually good and very close to FINDTAGS for the tag lines, but tag-jumping can occurs in the tracking process, due to the local nature of the method.
3. the mutual information based methods. This very recent group of methods works on the deformation itself, and maximizes mutual information [28] or

other similarity image information values [29]. It is computed with the grid image, but yields visually good results. No clear validation is available.

When comparing the different methods, the tracking of HARP is better since it is designed specifically for this application, but even HARP does not give acceptable results near the boundaries of the myocardium. For further applications, the useful part of the myocardium is thus limited.

2 Extrema Temporal Chaining (ETC)

Our strategy is mainly built on two elements. First of all, we want our method to be as robust and precise as possible, while being automatic and fast. The robustness demands a non-local approach. And we wish to have immediate post-processings once the tracking is done for the whole image. Both conditions can be fulfilled by matching points. We chose to achieve a geometric registration as it is the case in the first type of methods described above. These methods often propose a prior segmentation of the tag and/or of the myocardium. But segmenting shapes on MRI images (tagged or not) is a problem of very high complexity, often requiring user interaction [9,11,14]. Moreover, myocardium contours are badly defined due to the movement of the heart through the imaging plane. However, the 2D-displacement field estimation may be done without any segmentation. Therefore, to obtain an automatic computation, we choose a geometric matching of tags through time rather than their direct spatial localization. Next 4 sections describe how this matching is done.

Secondly, tagged data consist of a set of two sequences of images with orthogonal tags as presented in Fig. 1. Rather than working on a grid image where both directions are mixed (and where signal information is lost), each direction is treated separately. Then, the results obtained in each case are combined.

2.1 Minima Chaining

One advantage of the CSPAMM tagging technique is that it produces very thin tags (1 pixel wide). They correspond to local directional minima, when considering the orthogonal direction of the tag lines. If one extracts the 1D-signals in this direction (that is orthogonally to the tags), then local directional minima become local minima. This is illustrated in Fig. 2.

As the heart never rotates more than 45 degrees, this extraction is made for the entire sequence, and the intersection of each persistent tag line with the extracted fixed line still corresponds to a local minimum. If tagging in the main directions of the image, the 1D-signal extraction simply corresponds to a row and column extraction.

We notice that the corresponding minima positions are very precise and close to each other between two consecutive time frames. So we achieve a temporal chaining of those minima (see Fig.2) in a very simple manner. All the minima are extracted at frame 1. They define the origins of the chains. Throughout

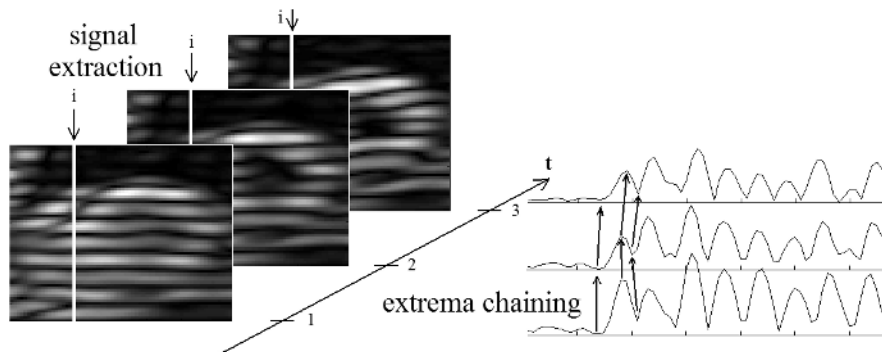


Fig. 2. Signal extraction and extrema chaining. Successive 1D-signals (right) are located at a fixed position in the tagged sequence. The extrema temporal chaining is represented with arrows.

the time, two minima on two successive images are linked if they have close spatial positions : if no minimum is found at frame $t + 1$ at the location where a chain ends at frame t , then the search extends to a neighborhood of size 1, 2, then 3 pixels. Each pixel is chained only once. This guarantees that there is no crossing between the different matchings. If one pixel at frame $t + 1$ may be associated with two different chains with the same shift, then the chain without other possible extremity is chosen. Otherwise the choice is made on a contrast criterion as defined in Sec. 2.3. At the end of the chaining procedure, every chain terminating before the end of the sequence is removed : this provides a kind of robustness by ensuring that remaining chains have a spatio-temporal meaning.

According to the tag orientation, rows and columns are successively considered. Results obtained by this temporal chaining procedure are promising : a good detection of the tag lines through the myocardium is achieved (see Fig. 3). There still remain some chains in the background, but they do not disturb further processing.

2.2 Spatial Smoothness Assumption of Displacements

As presented on Fig. 3, a few matching errors occur, though the false detections are isolated in space and time. These errors may easily be corrected by adding a spatial smoothness assumption of the deformation field, which comes from the elasticity of the myocardium [1,32]. To incorporate this, instead of choosing the closest minima like in section 2.1, we shift the prospection according to neighboring predicted displacements. Considering the lines sequentially, we may predict displacements on line number n from frame t to $t + 1$ using chains built up at the preceding steps (lines $n - 1$ to $n - 3$). This is possible because when handling line n , chains on the 3 previous lines have been treated and their position at frame $t + 1$ is known. Thus we may compute their coarse displacements from t

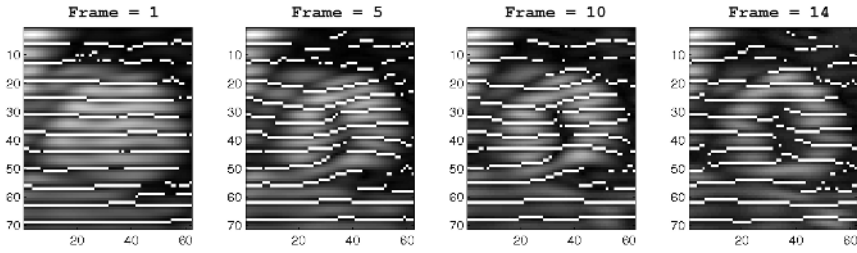


Fig. 3. Vertical-temporally chained minima on frames 1, 5, 10 and 14 in white, resulting from the columns extraction of a horizontal-tagged sequence

to $t + 1$ via a linear interpolation from the minima locations. This is computed in parallel with the minima chaining and used to shift the expected locations of chains.

2.3 Proximity to Myocardial Boundaries

Let us now focus on what happens near the boundaries of the myocardium. Due to the tissue properties and to the through-plane motion of the heart, the boundaries are much noisier than the inner-wall. Consequently, those points are critical locations where motion estimation is often least reliable.

We want our minima chainings to stick to the myocardium. However, near the boundaries the chain might be lost in the noisy background which presents many irrelevant minima. The critical situation that has to be avoided is when a minimum just beside the myocardium jumps to a background minimum. In such cases, there is a huge loss of "dynamics", which can be easily detected. We choose a symmetrical definition of the dynamics of a minimum by the formula

$$\frac{M_{+1} + M_{-1} - 2 \times m}{2}$$

where M_{+1} , M_{-1} are the luminous intensity values of the 2 surroundings maxima and where m is the value of the minimum for which we calculate the dynamics.

How is the dynamics criterion incorporated in our procedure? Instead of selecting only one minimum as in section 2.2, the two best candidates are selected. The second is preferred if its dynamics is twice as great as the first one. A coefficient value of 2 avoids a possible confusion between two inner tags.

2.4 Including Maxima and Sub-pixel Accuracy

For instance, only the minima of lines are chained. A similar minima chaining is obtained when considering the horizontal tags and the image columns. Furthermore, the same procedure may be applied on the maxima since the problem is dual. Tags being fine lines, the minima chaining is accurate. White areas being

wider, false chaining may occur : maxima chains may cross minima chains. In that case, the maxima chain is removed and the procedure is continued (the maxima chaining is supposed to be less robust than the minima chaining). At the end of the chaining procedure, we have four sets of chains corresponding to two directions of tags and two types of extrema (minima and maxima). The displacement field can now be estimated using all this information and not only the minima chaining.

Before computing the 2D-displacement field, the position of each extremum is refined by a polynomial interpolation using the 2 surrounding pixels. We obtain a sub-pixel accuracy and remove the discontinuity between "no displacement" and a 1-pixel displacement. However, the sparse displacement information directly available from those chainings corresponds to an apparent displacement of the tag lines along their orthogonal direction. It is not a component of the real displacement : similarly to the through-plane effect, the signal extraction step creates a through-line effect due to the motion of the heart across the extracted line.

There still remain two tasks to accomplish : retrieve the real components of the 2D-displacements of extrema, and interpolate those sparse measurements. Next section is devoted to these goals.

3 2D Deformation Field Computation

3.1 2D Displacements Real Components Retrieval

We know that when a point of the image at frame t lies on a tag (resp. crest) line, its material matching lies on the same tag (or crest) line at frame $t + 1$. This is used to retrieve the real 2D displacements components. For the moment, the extrema (the minima as well as the maxima) have been chained through time. We now need to chain them spatially in order to define the tags and the crest lines. Thanks to the temporal chaining, a unique spatial chaining may be conserved for all the frames. There is indeed a one-to-one association of the extracted points of the tags. Furthermore, the spatial chaining may be very easily and robustly computed on the first image of the sequence since for this image tags are less deformed and quite close to lines and columns of the image.

Now, the computation of the real 2D-displacements requires a correction of the trajectory of the chained extrema (Fig. 4). The target point is shifted from its initial to another extracted line, according to its orthogonal displacement, following the tag or crest line. In this manner, the displacement map of all horizontal chains points is updated using the vertical displacement components and conversely for all the vertical chains using the horizontal displacements. One component of the 2D-displacement being used to update the other one, the procedure is not symmetrical and hence must be iterated. In practice, due to the low values of the components, two or three iterations are sufficient.

3.2 Interpolation Using the Laplacian Equation

Obviously, the myocardium motion being continuous, the interpolation scheme has to respect some smoothness conditions. We work on each component of the

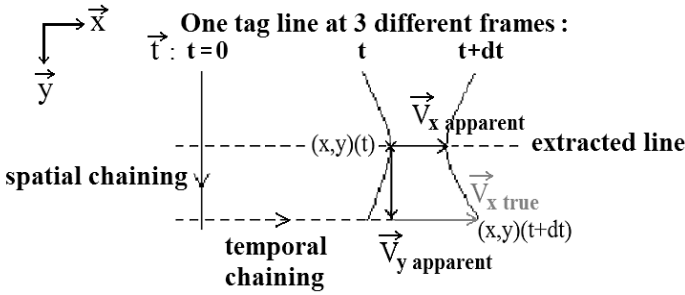


Fig. 4. True and apparent V_x components on a horizontal chain. Vertical lines along y represent the same tag line at frames 0, t , and $t + dt$, of which points have been temporally chained on the dashed extracted lines. The gray vector V_{xtrue} is closer to the real x -component of the 2D-displacement.

vector field separately since their continuity will induce the continuity of the entire vector field.

In order to prevent the apparition of artificial velocity values, we reject methods like splines or polynomial approximations, since they create extrema. According to the locations where displacement values are well-known, a bilinear interpolation is not well-adapted. Therefore, we use an iterative interpolation according to the Laplacian equation with Dirichlet limit conditions¹ :

$$\Delta D_x(x, y) = 0 \text{ et } \Delta D_y(x, y) = 0$$

where (D_x, D_y) is the displacement vector field. It has 3 advantages over other methods. Firstly, it does not create any new extremum. Secondly, it does not propagate erroneous values : they are stopped by the surrounding conditioning points, and induce a Dirac-like result. And thirdly it is very easy to perform, since it is the limit state of a recursive mean scheme. Thanks to this interpolation behavior, erroneous values remain do not propagate. Hence we can detect them very easily by computing and thresholding the gradient. Finally the reverse displacements are interpolated in the same way.

4 Results and Validation

4.1 Results

As presented in the two preceding sections, we dispose of a fully automatic procedure for computing the dense 2D displacement field from a couple of tagged MRI sequences. It has to be pointed out that the proposed Extrema Temporal Chaining (ETC) method does not rely on the fading effect and does not use any physical value nor any variable parameter. The ETC algorithm was implemented using Matlab7.0, the MathWork, Inc. script on a Dell Latitude, 1.39 GHz, 1.00

¹ Δ designs the Laplacian operator.

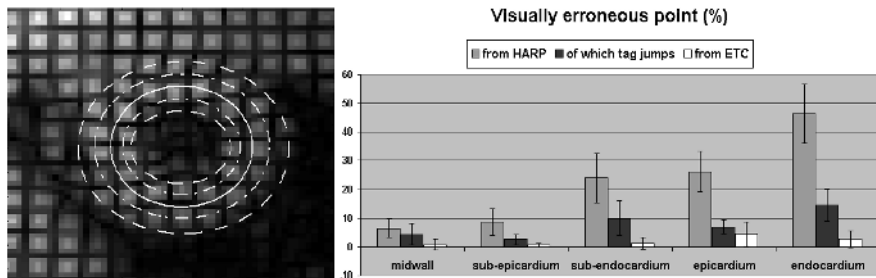


Fig. 5. Visually unacceptable points (%) tracked with HARP and ETC methods. The different zones are defined by curves set on the midwall (plain), sub-endocardium, and sub-epicardium (dash-dotted), endocardium, and epicardium (dashed). The percentage of obvious false matching are given for HARP, of which tag jumps, and for ETC. The results are based on 16 sequences from 3 rats and 27 sequences from 5 patient studies.

Go, pentium 4. The mean computing time varies from 20 seconds for rat hearts (size around 60×60 pixels, tag spacing of 4 pixels, and image delay of 13 ms) to 50 seconds for humans (size : 70×70 , tag spacing : around 7 pixels, image delay : 30 ms).

Another major advantage of our method may be noted : the post-processing is immediate. For example, to track any point within the myocardium, we only have to add its 2D displacements to its coordinates (with bilinear interpolation for sub-pixel accuracy). The results are visually good even near the boundaries, for rats and patients alike. But they are a bit more robust and more precise for humans than for rats, due to the limitation of acquisition parameters for imaging rat hearts (10 times smaller with a 5 times faster beating).

A limitation comes however from the acquisition parameters. A visually bad (blurred or noisy) sequence gives more doubtful results. Diminishing the tag spacing without sufficiently diminishing the temporal imaging delay may cause some errors such as tag jumping. But this compromise also has to be taken into account for other image processing methods. When working with adapted (and common) conditions, we can assume that the method works for every sequence.

4.2 Comparison to HARP

As a validation step, we compared ETC results with those of HARP tracking[27], because HARP happens to be a validated method which is largely used in research laboratories. HARP results are known to be good and accurate, except for some tag jumping occurring mainly near the boundaries. In addition to that phenomenon, some points near the boundaries may be lost in the background. HARP tracking of points on the myocardium is about 20 times longer than with our method. This is a useful advantage for ETC if we wish to track many points, such as needed in a grid visualization of the deformation for example.

We compared the tracking of both methods essentially on short axes (SA) sequences. We placed 5 curves near 5 locations : midwall, sub-epicardium, sub-endocardium, epicardium, and endocardium (see Fig.5, left image). The initial curves were set identical on the first frame, and the number of visual non-acceptable points was checked manually through the sequence for both methods. This means that we only rejected evident errors. For HARP, we distinguished classical tag jump errors from others like points lost in the background. This is recorded in Fig.5. The mean quadratic distance between HARP and ETC tracking at the final frame is about $\frac{1}{4}$ pixel in the midwall and $\frac{1}{2}$ pixel in the sub-epicardium and sub-endocardium areas when no visual error occurs. It raises to more than 1 and 2 pixels for epicardium and endocardium respectively (taking the erroneous points), corresponding to very irregular HARP outlines.

ETC appears to be more robust than HARP in each region, and the benefit is especially big for boundaries zones. Due to the imaging quality, rat heart series give somewhat worse but acceptable results.

5 Conclusion

The goal of this work was to propose a fast and automatic algorithm to compute the 2D apparent velocity field from two CSPAMM tagged MRI sequences. We ran the ETC algorithm on mainly 22 patients and 10 rats, accumulating 280 sequences. 200 of them concern patients, and are divided into 150 SA (Short Axes) views and 50 LA (long axes) views. The 80 others are composed of 70 SA and 10 LA of rat hearts. The results depend on the quality of the acquired cine, but are overall satisfactory. Moreover, ETC happens to have an accuracy close to HARP and to be much more robust on the boundaries. It might be used either alone to analyze tagged MRI, or as an intelligent initialization of target points for HARP tracking.

We are currently validating ETC results on every sequence with a blind marking of tag line crossings. The marking is done by 4 different people, including 3 clinicians. Next step is to include ETC to a medical imaging tool already widely in use at the GUH² for clinicians use.

We then intend to make a large use of it to compute meaningful suitable measures on the heart, such as strains, divergence, or movement modeling. More specifically, further research will attempt to better characterize abnormal zones of contraction resulting from infarction induced in rats.

References

1. R.S.Chadwick, Mechanics of left ventricle, Biophysical journal, 39:279-288, 1982
2. R.E.Henson et al., *Left ventricular torsion is equal in mice and humans*, Am J Physiol Heart Circ Physiol, 278:H1117-H1123, 2000
3. L.Axel, L.Dougherty, *MR imaging of motion with spatial modulation of magnetization*, Radiology, 171:841-845 1989

4. L.Axel, R.Goncalves, D.Bloomgarden *Regional heart wall motion : Two-dimensional analysis and functional imaging with MR imaging*, Radiology, 183:745-750, 1992
5. SE Fisher et al., *Improved myocardial Tagging Contrast*, Magnetic Resonance in Medicine, 30:191-200, 1993
6. D.J.Fisher, S.Collins, *Automated detection of noninvasive magnetic resonance imaging markers*, Computers in Cardiology, 493-496, 1991
7. L.Delman, J.T.Kent, K.V. Mardia, *Tracking of Tagged MR Images by Bayesian Analysis of a Network of Quads*, Proceedings of the 15th International Conference on Information Processing in Medical Imaging, 495-500, 1997
8. T.S.Denney Jr., *Estimation and detection of myocardial tags in MR images without user defined myocardial contours*, phd, electrical engineering department, auburn university, may 1997
9. M.A.Guttman, J.L. Prince, E.R.McVeigh, *Tag and contour detection in tagged MR images of the left ventricle*, IEEE Transactions on Medical Imaging, 13(1):74-88, March 1994
10. M.A.Guttman, E.A.Zerhouni and E.R.McVeigh, *Analysis and Visualization of Cardiac Function from MR Images*, IEEE, Computer Graphics and Applications, 17:30-38, 1997
11. S.Kumar and D.Goldgof. *Automatic tracking of CSPAMM grid and the estimation of deformation parameters from cardiac MR images*, IEEE Transactions on Medical Imaging, 13(1):122-132, 1994
12. A.A.Young and L.Axel. *Three-dimensional motion and deformation of the heart wall: estimation with spatial modulation of magnetization - a model-based approach*, Radiology, 185:241-247, 1992
13. A.A.Amini et al. *Energy-minimizing deformable grids for tracking tagged MR cardiac images*, Computers in Cardiology, 651-654, october 1992
14. A.A.Amini, R.W.Yasheng Chen Curwen, V. Mani, J.Sun, *Coupled B-snake grids and constrained thin-plate splines for analysis of 2-D tissue deformations from tagged MRI*, IEEE Transactions on Medical Imaging, 17:344-356, June 1998
15. A.A.Amini, Y.Chen, M.Elaiyadi, P.Radeva, *Tag Surface Reconstruction and Tracking of Myocardial Beads from CSPAMM-MRI with Parametric B-Spline Surfaces*, IEEE Transactions on Medical Imaging, 20(2), february 2001
16. H.-H.Chang, José M.F.Moura, Yijun L. Wu, K.Sato, C.Ho, *Reconstruction of 3D dense cardiac motion from tagged MR sequences*, In International Symposium on BioImaging, Crystal City, VA, April 2004.
17. S.K.Tadikonda, D.J.Fisher, S.M.Collins, *Automated detection of cine-CSPAMM magnetic resonance tags at sub-pixel resolutions*, Proceedings of Computers in Cardiology, 1992
18. B.K.P.Horn and B.G.Schunck, *Determining optical flow*, Artificial Intelligence, 17:185-203, 1981
19. M.A.Gennert and S.Negahdaripour, *Relaxing the brightness constancy assumption in computing optical flow*, M.I.T. Technical Report, June 1987. A.I. Memo No. 975
20. M.Kass, A.Witkin, and D.Terzopoulos, *Snakes: Active contour models*, Proceedings of International Conference on Computer Vision, 259-268, 1987
21. T.S.Denney Jr., *Stochastic estimation of deformable motion from magnetic resonance tagged cardiac images*, PhD thesis, Johns Hopkins University, 1994
22. J.L.Prince and E.R.McVeigh, *Motion estimation from tagged MR image sequences*, IEEE Transactions on Medical Imaging, 11(2):238-249, 1992
23. S.Gupta, *Optical Flow Techniques for Cardiac Motion Estimation from Tagged MR Images*, M.S.E. 1994, Ph.D. 1998

24. L.Dougherty, J.C.Asmuth, A.S.Blom, L.Axel, and R.Kumar, *Validation of an optical flow method for tag displacement estimation*, IEEE Transactions on Medical Imaging, 18(4):359-63, April 1999
25. S.N. Gupta, J.L. Prince, and S.Androutsellis-Theotokis, *Bandpass Optical Flow for Tagged MR Imaging*, Proceedings of International Conference on Image Processing, San Diego, pp. III:364-367, Oct. 1997
26. N.F.Osman and J.L Prince, *Angle Images for Measuring Heart Motion from Tagged MRI*, Proceedings of IEEE International Conference on Image Processing, vol.1, pp.704-708, Chicago, October 1998
27. N.F.Osman, *Measuring Regional Cardiac Function Using Harmonic Phase Magnetic Resonance Imaging*, (Advisor: Professor Prince), M.S.E. 1998, Ph.D. 2000
28. R.Chandrashekhara, R.H.Mohiaddin, and D.Rueckert, *Analysis of myocardial motion in tagged MR images using non-rigid image registration*, In Proc. SPIE Medical Imaging 2002: Image Processing, San Diego, CA, February 2002.
29. C.Petitjean, *Recalage non rigide d'images par approches variationnelles statistiques - Application l'analyse et la modélisation de la fonction myocardique en IRM*, Thesis / PhD (24 September 2003), ARTEMIS (Institut National des Télécommunications, Evry), UNIVERSITE RENE DESCARTES - PARIS V
30. W.S. Kerwin and J.L. Prince, *Cardiac material markers from tagged MR images*, Med. Image Anal., 2(4):339-353, 1998
31. W.G.O'Dell, C.C.Moore, W.C.Hunter, E.A.Zerhouni and E.R. McVeigh, *Three-dimensional myocardial deformations: calculation with displacement field fitting to tagged MR images*, Radiology, 195:829-835, 1995
32. T.S. Denney Jr., J.L. Prince, *Reconstruction of 3D Left Ventricular Motion from Planar Tagged Cardiac MR Images: an Estimation Theoretic Approach*, IEEE Transaction on Medical Imaging, 4(4):625-635, December 1995

Data Fusion and Fuzzy Spatial Relationships for Locating Deep Brain Stimulation Targets in Magnetic Resonance Images

Alice Villéger¹, Lemlih Ouchchane^{1,2}, Jean-Jacques Lemaire^{1,3},
and Jean-Yves Boire^{1,2}

¹ ERIM, Medicine Faculty of Clermont-Ferrand, Auvergne University, France
alice.villeger@u-clermont1.fr

² Biostatistics Unit, University Hospital of Clermont-Ferrand, France
lemlih.ouchchane@u-clermont1.fr, j-yves.boire@u-clermont1.fr

³ Department of Neurosurgery, University Hospital of Clermont-Ferrand, France
jjlemaire@chu-clermontferrand.fr

Abstract. Symptoms of Parkinson's disease can be relieved through Deep Brain Stimulation. This neurosurgical technique relies on high precision positioning of electrodes in specific areas of the basal ganglia and the thalamus. In order to identify these anatomical targets, which are located deep within the brain, we developed a semi-automated method of image analysis, based on data fusion. Information provided by both anatomical magnetic resonance images and expert knowledge is managed in a common possibilistic frame, using a fuzzy logic approach. More specifically, a graph-based *virtual atlas* modeling theoretical anatomical knowledge is matched to the image data from each patient, through a research algorithm (or *strategy*) which simultaneously computes an estimation of the location of every structures, thus assisting the neurosurgeon in defining the optimal target. The method was tested on 10 images, with promising results. Location and segmentation results were statistically assessed, opening perspectives for enhancements.

1 Introduction

Deep brain stimulation (DBS) is widely accepted to alleviate movement disorders. This neurosurgical technique is mainly used to treat severe idiopathic Parkinson's disease[1]. DBS is performed under stereotactic conditions (*i.e.* with a stereotactic frame fixed to the skull) and consists in the implantation of two electrodes (one per hemisphere) into an anatomical target. The electrodes are connected to a neuro-pacemaker, placed at the trunk, and used to control the stimulation parameters. This control device allows to manually adjust the stimulation parameters according to the reduction of antiparkinsonian drugs, in order to optimize clinical benefit against adverse effects. Various anatomical structures, located in the general area of the central gray nuclei, can be targeted:

e.g. the subthalamic nucleus (STN), the globus pallidus internus (GPi) and the ventral intermediate nucleus of the thalamus (Vim). The exact location of the optimal target remains unknown, as the biological mechanisms involved in DBS have yet to be fully understood. The chronic stimulation of these targets acts as a lesion (or ”-tomy”, as in ”pallidotomy”), but with a dramatic reduction of the drawbacks (mainly irreversibility, low control of lesion volumes, and severe adverse effects in case of bilateral lesions).

Beyond the clinical aspects, the implantation technique itself is still a matter of debate. Two main methods are proposed to define a surgical target. The classical stereotactic approach, historically based on ventriculography and stereotactic atlas [2][3], uses a proportional indirect method, *i.e.* the coordinates of the target are computed relatively to internal landmarks (the anterior and posterior white commissures, called AC/PC). A more recent approach, called direct targeting, relies on magnetic resonance imaging (MRI), which allows the direct visualization of the target[4][5][6]. Both indirect and direct methods having limitations (as the former struggles to take into account inter-individual variability, while the latter depends mainly on image quality), clinical protocols applied worldwide are still not standardized.

The clinical protocol routinely applied in our institution is based on direct targeting[7][8], and unfolds in two steps. The pre-implantation phase starts with the acquisition of tridimensional anatomical data under stereotactic conditions: namely, T₂ weighted MRI sequences ¹ resulting in three orthogonal anisotropic images (1 voxel = 0.52 × 0.52 × 2mm³) Using anatomical databases[2][3][9][10] as a reference, the neurosurgeon carries out a visual analysis of the images, in order to estimate the location of anatomical structures of interest, *i.e.* the stereotactic target and its surroundings. The manual labeling of the central gray nuclei area allows to determine an optimal implantation trajectory ² for the DBS electrodes. At implantation stage, the stereotactic system ³ is used as a fixed referential to insert the electrodes and reach the previously computed coordinates. Since surgery is performed under local anesthesia, the positioning of each electrode along the insertion axis can be optimized according to the clinical effects noticed during acute stimulation tests. Clinical results have been already reported [8][11].

The manual labeling step required at pre-operative stage is relatively time consuming, (about one hour for two sides) demands a high level of expertise, and (being operator dependent) raises reproducibility issues. In order to assist the practitioner in defining an optimal trajectory, we proposed to design a computer process for image analysis, able to automate the extraction of areas of interest (*i.e.* relevant anatomical structures such as the stereotactic targets and their neighborhood) from MRI. We focused on the STN as this target, which is the reference in DBS for Parkinson’s disease, is also difficult to identify due to its small size and its complex anatomical surroundings.

¹ Sonata 1.5 Tesla, Siemens, Germany.

² iPlan, BrainLab, Germany.

³ Leksell G frame, Elekta, Sweden.

2 Method

2.1 General Framework

It has been reported that recognition of brain structures in MRI could be achieved through a data fusion method based on fuzzy logic[12], where every sources of information (*i.e.* patient data and expert knowledge) are modeled in the same possibilistic frame. Indeed, fuzzy sets and the possibility theory framework appear suitable to account for the inter-individual variability observed in living tissues and anatomical structures. Particularly, encouraging preliminary results were obtained for the segmentation of the subthalamic nuclei[13].

A fuzzy membership map is an image mapping every voxel to a membership degree $\mu \in [0, 1]$, which quantifies how much the voxel belongs to a particular fuzzy set. Fuzzy tissue maps are extracted from the patient's MRI through a clustering step. Prior expert knowledge on the spatial relationships (*e.g.* relative distance or direction) between anatomical structures of interest (*i.e.* landmarks and the targeted structure itself) can also be expressed through fuzzy membership maps.

The complementary fuzzy maps are then fused, by means of suitable possibilistic operators, in order to achieve a segmentation of the targeted structures. The fusion process relies on the definition of a research route, or *scenario*: a pre-determined list of intermediate landmarks, linked by spatial relationships, which are to be segmented successively in order to reach the target.

This method requires the expert to devise a specific single route for each potential target. Furthermore, each intermediate structure has to be segmented accurately: otherwise, errors would be propagated along the research route, down to the final structure. Consequently, additional processings ensuring shape constraints (*e.g.* region growing, fuzzy shape maps, or edge detection) have to be introduced systematically, in order to refine the segmentation obtained by simply fusing the spatial relationships and tissue maps.

2.2 From *scenario* to *strategy*: Toward a Global Research Scheme

It appeared to us the fusion process could instead be guided entirely by a "virtual atlas", *i.e.* a model formalizing expert knowledge by describing every structure in relation to the others within a graph. The use of a relational graph to model prior anatomical knowledge has been suggested in other recent works, through either global or progressive approaches[14].

The global approach consists in matching a relational graph, extracted from an anatomical atlas, with a similar graph, extracted from previously segmented patient data. However, the over-segmentation of those images results in a difficult problem of inexact graph-matching. The progressive approach is based on a graph model containing both iconic (*i.e.* extracted from a digital atlas) and symbolic (*i.e.* expressed through linguistic descriptors) knowledge. Yet the associated recognition method still relies on a sequential research route which has to be predefined specifically for each targeted structure.

Instead of relying on predefined research sequences (or *scenario*), we aimed to design a recognition process (or *strategy*) leading to the simultaneous exploration of several research paths, resulting in the progressive segmentation of every structure defined in the model. While a research scenario depends on its target, such strategy, guided solely by prior expert knowledge, depends on the nature of the information contained in the graph-based model.

2.3 Graph-Based Model for Prior Expert Knowledge

The expert model is based on a graph, *i.e.* a list of vertices linked by edges. Each vertex represents an anatomical structure of interest (*i.e.* a target or landmark) and contains relevant information such as tissue composition. Each edge represents a spatial relationship between two structures, such as their relative directions. This generic model leaves much freedom concerning the exact nature of the data it contains: it is possible to add complementary types of information to both the vertices (*e.g.* morphology) or the edges (*e.g.* distance). A specific kind of information such as directions can also be represented relatively to points, or to whole objects, through various models.

These modeling choices partly depend on the available sources of information. Prior knowledge can be provided directly by the expert, expressing qualitative information by means of semantic descriptors (*e.g.* "above" or "anterior to"). Such formal descriptions are often imprecise and unexhaustive, but supposedly reliable and relevant. Conversely, precise quantitative information on every structures and relationships can be extracted from a set of pre-labeled images, but the relevance and statistical representativity of the data remains uncertain.

In our feasibility study, we chose to focus on reliable information which could be obtained either from the expert or from images: tissue composition and relative direction. Actually, direction is a complex information, involving angle and distance relationships, as well as the shape of the objects. We settled for a simple representation method requiring little prior knowledge, and taking little time to compute. Structures are assimilated to their center of gravity, and directional relationships between these single points are defined according to their projections along three orthogonal axis: "above/below", "anterior/posterior" and "left/right".

2.4 Research Strategy

We designed a simple automated fusion algorithm which uses the information contained in the virtual atlas to guide the whole segmentation process. This particular method takes advantage of a specific property of our model: directional relationships between lined up points translate into partial order relations, allowing us to sort structures along each main axis.

The very first step consists in a tissue classification[15] of the patient's MRI. Each structure's membership map is then initialized with the fuzzy map of the tissue it belongs to. Throughout the whole process, these membership maps are used to estimate a structure's location by computing its center of gravity.

The next step of the algorithm consists in roughly separating structures belonging to the same tissue class. Spatial relationships are propagated from the most central structures (as the initial location of their center of gravity is assumed to be closer to the solution) toward the most outer ones, then back from the border toward the center (in order to constrain the central structures). "Central" and "outer" refer to the relative position of the structure's projection on a given axis. This whole separation step relies on the hypothesis that the structures' repartition is quite homogeneous, and that the image contains every structure described in the model: discrimination between structures of the same tissue is based on their relative spatial characteristics.

The final step of the process consists in an iterative refinement of that first solution, in order to ensure that every spatial constraint is respected: all relationships in the model are propagated simultaneously and iteratively, until convergence toward a stable solution is achieved (i.e. when the difference between the maps computed at step n , and the maps computed at step $n+1$, is close to zero). The fuzzy membership map obtained for every structure can then be interpreted as crisp segmentation results, once a defuzzification threshold (set to 0.5 by default) has been applied: membership degrees are set to 1 when above the threshold, and set to 0 when below.

Propagating a relationship from a S_1 structure toward a S_2 structure means redefining the membership map of the S_2 structure by fusing it with the directional map obtained from the S_1 structure. But there is no unique method for computing a fuzzy directional map, even from a single point. We used a simple function projecting coordinates along the direction axis, then setting the membership degree to 1 when the directional constraint was respected (for instance, $y_1 > y_2$ when the models said " S_1 is above S_2 "), or to 0 when it was not.

3 Results

3.1 Test Protocol

First preliminary tests were performed on simulated data, *i.e.* randomly generated 3D images on which structures were represented by spheres of various random radius. The spheres were colored in various gray scales (with Gaussian noise added) in order to simulate structures belonging to a specific tissue class. A graph model was generated along with the simulated image, then used to segment it. Right after the first step of the algorithm, each structure was already correctly identified: the computed center of gravity was located within the corresponding sphere, though small segmentation errors (underestimation) could occur in case of close structures.

More realistic tests were performed afterward on a sample of ten images from parkinsonian patients. The MRI sets were composed of $512 \times 512 \times 24$ voxels of $0.52 \times 0.52 \times 2\text{mm}^3$. An appropriate coronal slice containing the STN was selected for each patient and pre-labeled by the expert for further comparison. We used a 2D anatomical model containing seven structures of interest. Every structure was allocated to a specific tissue class among the three taken into account, *i.e.*

#	Structure	Tissue
1	Lateral Ventricle	CSF
2	Third Ventricle	CSF
3	Thalamus	GM
4	Caudate Nucleus	GM
5	Pyramidal Tract	WM
6	Lenticular Nucleus	GM
7	STN + Substantia Nigra	GM

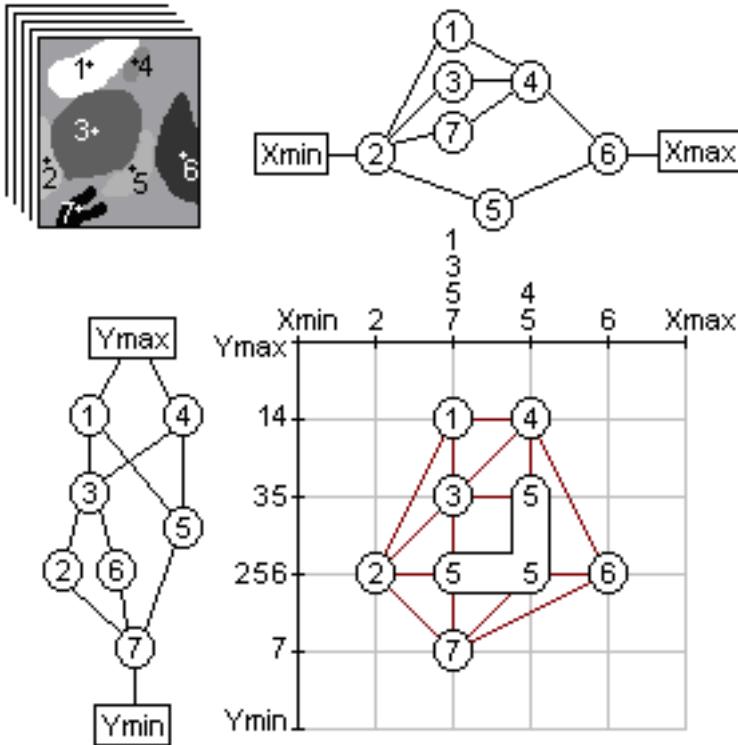


Fig. 1. Anatomical model

cerebrospinal fluid (CSF), white matter (WM) and gray matter (GM). Direction relationships were defined based on directional invariants observed on a few pre-labeled images (classical anatomical atlases, as well as four patients not used in the evaluation ; *c.f.* Figure 1) and validated by the expert. The model, defined for the right hemisphere, can easily be adapted for the left hemisphere by reversing the "left/right" relationships.

A rectangular Region of Interest (ROI) containing every structure defined in the model was defined manually in order to constrain the fusion algorithm to that particular region. This was the only non-automated step of the process. Preliminary tissue clustering was performed on each slice, resulting in four tissue

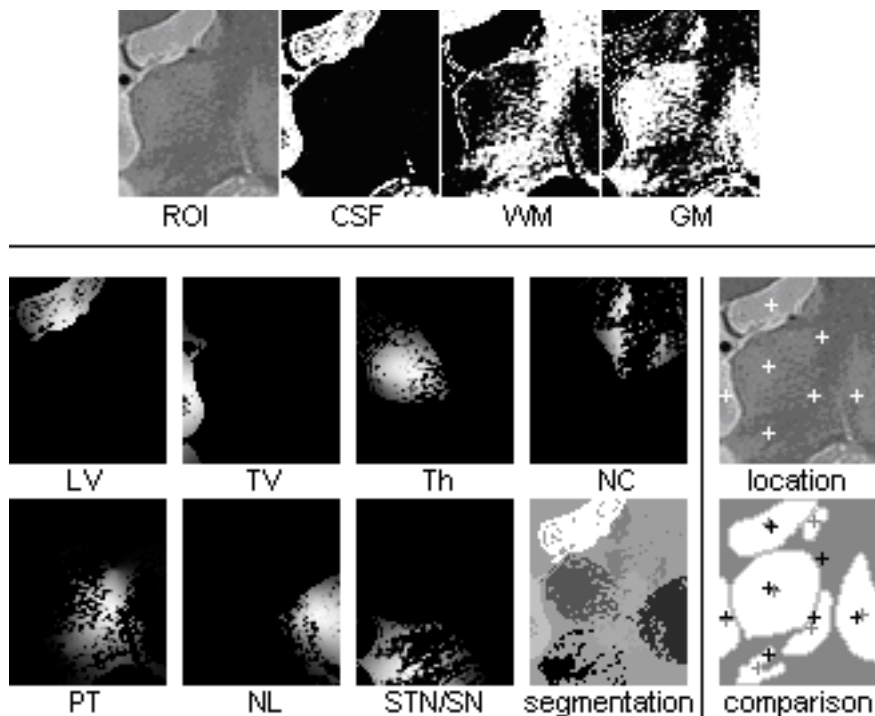


Fig. 2. Tissue classification, followed by segmentation and location results. The first image on the top, labeled "ROI", refers to the original region of interest. The 3 other images are the tissue membership maps obtained through the classification step, and used to initialize our algorithm. The bottom part features segmentation and location results. Seven of the images on the left part refer to the fuzzy membership maps obtained for every structure in the model. The 8th image, labeled "segmentation", represents the defuzzified segmentation results. On the right part, the image labeled "location" features the computed center of gravity for every structure. The last image, labeled "comparison", represents the expert's segmentation, along with the corresponding centres of gravity (gray crosses), while the black crosses refer to the centres of gravity computed with our method (also featured on the "location" picture).

maps: CSF, WM, GM and background (not used in our model). The first three tissue maps (cropped to fit the ROI) were then used to initialize the fusion algorithm. For each image, the final step of the fusion algorithm converged after two or three iterations. The whole process (tissue clustering and fusion algorithm) took about 30 seconds. Figure 2 illustrates results obtained for one patient.

The expert assessed the location results as satisfactory. Every structure, except for the CN, was correctly located: when compared with the expert's labeled image, the computed center of gravity was always located within the corresponding structure. However, the membership maps contain holes and unconnected

Table 1. Statistical analysis of the results on every patients, for every structures

Structure	ICC on X	ICC on Y	average TC	standard deviation on TC
LV	.89	.99	.67	.11
TV	.97	.95	.34	.16
TH	.96	.98	.33	.14
CN	.49	.91	.15	.13
PT	.89	.94	.24	.05
NL	.93	.97	.31	.06
STN	.74	.98	.27	.06
Mean value	.84	.96	.33	.10

parts, and some of them tend to differ significantly from the expert’s labeling: small structures such as the STN/SN group tend to be overestimated, at the expense of larger structures such as the thalamus, which seems underestimated.

3.2 Statistical Analysis

For every structure, over the ten patients, results have been statistically analyzed in term of location and segmentation, by comparing them with the results obtained with the reference method, *i.e.* expert labeling. Location results (the coordinates of the center of gravity) were assessed using intraclass correlation coefficients (ICC) between the computed values and the reference values, while segmentation results were assessed using the Tanimoto coefficient[16]:

$$TC(S_{result}, S_{reference}) = \frac{|S_{result} \cap S_{reference}|}{|S_{result} \cup S_{reference}|}$$

The impact of the defuzzification threshold has also been studied.

The statistical analysis confirmed our first visual estimations (Table 1). The ICC, with an overall average value of 90%, showed almost perfect concordance for the location of most structures, the only notable exception being the abscissa of the caudate nucleus (49%). On the other hand, the TC, with an average value of 33%, emphasized the mediocre overall quality of the segmentation, which would differ greatly among structures (from 15% for the caudate nuclei to 67% for the lateral ventricule).

Our tests also revealed some structures exhibited significantly different reactions to the setting of the threshold parameter. Two groups could be distinguished: the LV, Thalamus, LN, and PT (or "large structures"), on the one hand, and the TV, CN, and STN/SN (or "small structures"), on the other hand. The large structures reacted favorably to a very low threshold value, while the small structures required a higher one.

4 Discussion

The statistical concordance of the location results is encouraging, especially considering the mediocre quality of the segmentation itself: it proves the structures

do not have to be perfectly segmented in order to be properly located. The inaccuracies observed in the segmentation are not surprising: holes and unconnected parts were to be expected since the model did not define any morphologic constraints. Nevertheless, naturally connected structures belonging to the same tissue class, such as the lateral ventricle and the third ventricle, could still be correctly set apart. This result demonstrates that directional relationships alone can be powerful tools, when used concurrently, for guiding the identification and segmentation of anatomical structures. In order to refine the initial segmentation (and correct small tissue clustering errors in the process), the current results may be used to initialize a method of competitive region growing or edge detection. Preliminary tests involving connected component labeling have already shown promising results.

Errors in the segmentation (and location, in the particular case of the caudate nucleus) seem mostly due to a size difference between the various structures. As the simple model used in this preliminary study represents structures by their center of gravity, with no complementary information on their relative size, all the resulting segmented structures tend to be roughly of the same size (this is also a consequence of the separation algorithm used, which assumes an homogeneous repartition of these centres of gravity). Consequently, large structures are often underestimated while small structure are overestimated. This is why we chose to consider the subthalamic nucleus and substantia nigra (which are both relatively small, and very close) as a single structure. This result emphasizes the need for a multiscale approach. The scale of the model could be adjusted to take into account different layers of details: a set of small neighboring structures could be gathered to define a super-vertex, while a single huge structure could be split into several sub-vertices. The resulting inclusion relationships between these new vertices could be expressed quite easily by simply adding corresponding inclusion edges to the graph. The inclusion edges would form a tree between the various layers of detail, resulting in a multiscale atlas. Of course the whole process would have to be adjusted to factor in the various scales. For instance, the fusion algorithm could be used successively for each layer.

The method should also be evaluated on real 3D images. Actually, preliminary tests have already been performed, showing results similar to those observed on the 2D ROI, but which could not be formally assessed as the expert had not yet labeled the area. Moreover, the low depth resolution of the images used in the clinical protocol (*c.f.* 3.1) is a source of imprecision in the location along the z axis. This is why we are currently working on a process for fusing the complementary information provided by the three complementary perpendicular anisotropic images, into a single isotropic image. Such a 3D reconstruction at a higher resolution implies solving an inverse problem similar to the one involved in computed tomography.

Concerning the future clinical validation of the process, we obtained a high-resolution *post mortem* MRI (4.7 Tesla) from a histologic slice[10]. This data will be used as a reference for evaluating segmentation results computed from a

standard MRI acquired from the same individual. The neurosurgical department of our institution has also granted us access to a large database of patients who have already been treated through DBS and can be studied *a posteriori*.

Once perfected and validated, our method should relieve the physician from the time-consuming process of manual labeling, providing him with useful assistance for the positioning of electrodes. By precisely identifying the optimal area for DBS, it could lead to a better understanding of the clinical phenomena involved. In the long run, the approach might also be extended to guide robotic surgery.

References

1. Walter, B., Vitek, J.: Surgical treatment for parkinsons disease. *Lancet Neurol.* **3** (2004) 719–728
2. Talairach, J., David, M., Tournoux, P., Corredor, H., Kvasina, T.: Atlas d'anatomie stéréotaxique. Masson et Cie, Paris (1957)
3. Schaltenbrand, G., Bailey, P.: Introduction to stereotaxis with an atlas of the human brain. Volume 2. Georg Thieme Verlag, New York NY, Stuttgart (1959)
4. Lemaire, J.J., Durif, F., Boire, J.Y., Debilly, B., Irthum, B., Chazal, J.: Direct stereotactic mri location in the globus pallidus for chronic stimulation in parkinson's disease. In: *Acta Neurochir (Wien)*. Volume 141. (1999) 759–766
5. Coubes, P., Vayssiere, N., Fertit, H.E., Hemm, S., Cif, L., Kienlen, J., Bonafe, A., Frerebeau, P.: Deep brain stimulation for dystonia: Surgical technique. In: *Stereotact Funct Neurosurg*. Volume 78. (2002) 183–191
6. Bejjani, B., Dormont, D., Pidoux, B., Yelnik, J., Damier, P., Arnulf, I., Bonnet, A.M., Marsault, C., Agid, Y., Philippon, J., Cornu, P.: Bilateral subthalamic stimulation for parkinson's disease by using three-dimensional stereotactic mri and electrophysiological guidance. *J Neurosurg* **92** (2000) 615–25
7. Lemaire, J., Durif, F., Debilly, B., et al.: Deep brain stimulation in the subthalamic area for severe idiopathic parkinson's disease: location of plots in the peroperative phase and at the three month follow-up. In: *Parkinson's and related disorders*. Volume 7. (2001)
8. Caire, F., Derost, P., Coste, J., Bonny, J.M., Durif, F., Frenoux, E., Villéger, A., Lemaire, J.J.: Stimulation sous-thalamique dans la maladie de parkinson sévère : étude de la localisation des ontacts effectifs. *Neurochirurgie* (2005) *in press*.
9. Parent, A.: Basal Ganglia. In: *Carpenter's human neuroanatomy*. Williams and Wilkins, Baltimore (1996)
10. Lemaire, J.J., Caire, F., Bonny, J.M., Kemeny, J., Villéger, A., Chazal, J.: Contribution of 4.7 tesla mri in the analysis of the mri anatomy of the human subthalamic area. In: *Acta Neurochirurgica*. Volume 8. (2004) 906–907
11. Durif, F., Lemaire, J.J., Debilly, B., Dordain, G.: Acute and chronic effects of antero-medial globus pallidus stimulation in parkinson's disease. In: *J Neurol Neuro Psy*. Volume 67. (1999) 315–21
12. Barra, V.: Fusion d'images 3d du cerveau : étude de modèles et applications. In: *PhD Thesis (Université d'Auvergne, Clermont-Ferrand)*. (2000)
13. Barra, V., Lemaire, J.J., Durif, F., Boire, J.Y.: Segmentation of the subthalamic nucleus in mr images using information fusion - a preliminary study for a computer-aided surgery of parkinson's disease. In: *MICCAI*. (2001)

14. Bloch, I., Colliot, O., Camara, O., Géraud, T.: Fusion of spatial relationships for guiding recognition, example of brain structure recognition in 3d mri. *Pattern Recognition Letters* **26** (2005) 449–457
15. Barra, V., Boire, J.Y.: Tissue segmentation on mr images of the brain by possibilistic clustering on a 3d wavelet representation. In: *Journal of Magnetic Resonance Imaging*. Volume 11. (2000) 267–278
16. Tanimoto, T.: Ibm internal report. Technical report, IBM (1957)

Robust Tracking of Migrating Cells Using Four-Color Level Set Segmentation*

Sumit K. Nath, Filiz Bunyak, and Kannappan Palaniappan

MCVL, Department of Computer Science, University of Missouri-Columbia, MO, USA
{naths, bunyak, palaniappank}@missouri.edu

Abstract. Understanding behavior of migrating cells is becoming an emerging research area with many important applications. Segmentation and tracking constitute vital steps of this research. In this paper, we present an automated cell segmentation and tracking system designed to study migration of cells imaged with a phase contrast microscope. For segmentation the system uses active contour level set methods with a novel extension that efficiently prevents false-merge problem. Tracking is done by resolving frame to frame correspondences between multiple cells using a multi-distance, multi-hypothesis algorithm. Cells that move into the field-of-view, arise from cell division or disappear due to apoptosis are reliably segmented and tracked by the system. Robust tracking of cells, imaged with a phase contrast microscope is a challenging problem due to difficulties in segmenting dense clusters of cells. As cells being imaged have vague borders, close neighboring cells may appear to merge. These false-merges lead to incorrect trajectories being generated during the tracking process. Current level-set based approaches to solve the false-merge problem require a unique level set per object (the N-level set paradigm). The proposed approach uses evidence from previous frames and graph coloring principles and solves the same problem with only four level sets for any arbitrary number of similar objects, like cells.

1 Introduction

Understanding behavior of migrating cells is becoming an emerging research area with many important applications. Behavior of migrating cells are important parameters of interest in understanding basic biological processes such as tissue repair, metastatic potential, chemotaxis, differentiation or analyzing the performance of drugs. Accurate segmentation and tracking of cells are vital steps in any cell behavior study.

In this paper, we present an automated cell segmentation and tracking system designed to study migration of cells imaged with a phase contrast microscope. Segmentation is performed using active contour level set methods with a novel extension that efficiently prevents false-merge problem. Tracking is done by resolving frame to frame correspondences between multiple cells using a multi-distance, multi-hypothesis algorithm. Cells that move into the field-of-view, arise from cell division or disappear due to apoptosis are reliably segmented and tracked by the system.

Simultaneous tracking of multiple cells imaged with a phase contrast microscope is a challenging problem due to difficulties in segmenting dense clusters of cells. As cells

* This work was supported by a U.S National Institute of Health NIBIB award R33 EB00573.

being imaged have vague borders, close neighboring cells may appear to merge. These *false-merges* lead to incorrect trajectories being generated during the tracking process. Other challenges for tracking include high number of cells, non-linear motion, lack of discriminating features, mitosis (cell division), and fragmentation during segmentation. In [1], Chan and Vese presented an algorithm to automatically segment an image $I(\mathbf{y})$ into *two* distinct regions (or phases) by minimizing a minimal partition Mumford-Shah functional. A multiphase variant of the same algorithm was also proposed to handle 2^n unique phases [2]. However, as observed by Zhang *et al.*, [3], Dufour *et al.*, [4] and, Zimmer and Olivo-Marin [5], the two variants of the Chan and Vese algorithm are unsuitable for reliable cell segmentation due to the problem of apparent merges in cells.

Zhang *et al.*, proposed a N -level set framework with an implicit coupling constraint to reliably segment cells in an image sequence [3]. While this alleviates the problem of apparent merging of cells, it is computationally expensive to implement. The approach we propose uses evidence from previous frames and graph coloring principles and solves the same problem with only four level sets for any arbitrary number of similar objects.

The organization of this paper is as follows. Section 2 describes the segmentation module. Salient features of our four-color level set segmentation algorithm are presented along with the related work, variants of the Chan and Vese level set algorithms and N -level set variant of Zhang *et al.*, algorithm [3]. Section 3 describes the tracking module. Comparative results and a discussion are presented in Section 4, while a conclusion is presented in Section 5.

2 Cell Segmentation Using Active Contour Level Set Methods

Accurate segmentation of individual cells is a crucial step in robust tracking of migrating cells as both over-segmentation (fragmentation) and under-segmentation (cell clumping) produce tracking errors (i.e., spurious or missing trajectories and, incorrect split and merge events). In this section, we describe three different level set segmentation methods and compare their performance for separating closely adjacent and touching cells in a dense population of migrating cells. The three techniques described in this section are all based on the “active contour without edges” energy functional with appropriate extensions, and include multi-phase Chan and Vese [2] (CV2LS), N -level sets with energy-based coupling by Zhang *et al.*, [3] (ZZNLS), and our novel four-color level sets with energy-based and explicit topological coupling [6] (NBP4LS-ETC). The latter two techniques use coupling constraints in order to prevent the merging of adjacent cells when they approach or touch each other.

- CV2LS: In order to segment multiple (i.e., N distinct) objects, Vese and Chan extended their previous 2-phase algorithm [1] by using $\lceil \log_2 N \rceil$ level sets [2]. The corresponding energy functional $E_{pc}(\mathbf{c}, \Phi)$

$$E_{pc}(\mathbf{c}, \Phi) = \sum_{1 \leq i \leq N} \mu_i \int_{\Omega(\mathbf{y})} (\mathbf{I}(\mathbf{y}) - c_i)^2 \chi_i d\mathbf{y} + \sum_{1 \leq i \leq \lceil \log_2 N \rceil} \nu_i \int_{\Omega(\mathbf{y})} |\nabla H(\phi_i)| d\mathbf{y}$$

where, N is the number of phases (i.e., regions in the image) associated with $\lceil \log_2 N \rceil$ level set functions, \mathbf{I} is the gray-level image being segmented, Φ is a

vector of level set functions, \mathbf{c} is a vector of mean gray-level values (i.e., $c_i = \text{mean}(\mathbf{I})$ in the class i), χ_i is the characteristic function for each class i formed by associated Heaviside functions $H(\phi_i)$, and (μ_i, ν_i) are constants associated with the energy and length terms of the functional, respectively.

The $\lceil \log_2 N \rceil$ level set formulation improves on the performance of a single level set function, as more number of objects with varying intensities can be efficiently classified. But does not prevent under-segmentation (i.e. incorrect merges), when the objects have similar intensities (i.e. cells).

- ZZNLs: To overcome the drawbacks of classical Chan and Vese level set formulations, while at the same time solving the problem of apparent merging of cells during tracking, Zhang *et al.*, [3] proposed a new model using N -level sets for segmenting cells. Here N is the number of cells at a given time instance. An *a priori* knowledge that cells do not merge during the evolution process was used to guide the segmentation process. This was achieved by a pair-wise energy-based coupling constraint on the level sets evolution process. A similar formulation was used by Dufor *et al.*, in 3D cell segmentation [4].

The energy functional, $E_{nls}(\mathbf{c}_{in}, c_{out}, \Phi)$, used to solve the evolution of N -level sets is given by [3]:

$$\begin{aligned}
 E_{nls}(\mathbf{c}_{in}, c_{out}, \Phi) = & \gamma \sum_{i=1}^N \sum_{j=i+1}^N \int_{\Omega} H(\phi_i)H(\phi_j) \, d\mathbf{y} + \nu \sum_{i=1}^N \int_{\Omega} |\nabla H(\phi_i)| \, d\mathbf{y} \\
 & + \mu_{in} \sum_{i=1}^N \int_{\Omega} (I - c_{in}^i)^2 H(\phi_i) \, d\mathbf{y} + \mu_{out} \int_{\Omega} (I - c_{out})^2 \prod_{\substack{i=1 \\ \forall i: H(\phi_i) < 0}}^N (1 - H(\phi_i)) \, d\mathbf{y}
 \end{aligned}
 \tag{1}$$

Here, $\Phi = [\phi_{i: i=1\dots N}]$ represents N -level sets associated with N cells in the image; \mathbf{c}_{in} represents average intensities of cells for $H(\phi_i) \geq 0$ while c_{out} is the average intensity of the background¹. The first term of the functional penalizes pair-wise couplings between level sets, while the second term controls the length of ϕ_i . $\mu_{in}, \mu_{out}, \gamma, \nu$ are constants associated with the functional.

- NBP4LS-ETC: The N -level set formulation described previously is able to overcome the apparent merging of neighboring cells. However this approach is not very scalable and is computationally expensive since for N objects it requires $N^2/2$ couplings. To overcome the computational cost, while still preventing the incorrect merges, we propose an *optimized* version of the N -level set algorithm described previously.

Our optimization is based on the fact that only neighboring cells can potentially merge. Through Delaunay triangulation the cell-to-cell neighborhood relationships are identified and represented in a graph where vertices represent the cells and edges represent the neighborhood relations.

The four-color theorem [7, 8, 9] states that any planar graph is four-colorable such that no two neighboring vertices have the same color. Thus, four rather than

¹ The region *exterior to all level sets* indicates the background.

N – level sets would suffice to classify N –objects (i.e., cells) in an image while insuring that neighboring objects do not share the same level set.

In order to evolve the four level sets we propose minimizing an energy functional, $E_{fc}(\mathbf{c}_{in}, \mathbf{c}_{out}, \Phi)$, shown in Eq. 2. The first two terms of the right-hand side of Eq. 2 are used to compute average intensities ($\mathbf{c}_{in}, \mathbf{c}_{out}$) within each level set, and outside all level sets, respectively. Using an *a priori* assumption that all the foreground objects (i.e., cells) in the image have very similar characteristics, we use a single average intensity c_{in} (i.e., $\forall i, c_{in}^i = c_{in}$). Whereas Zhang *et al.*, model of computing average intensities for each cell (Eq. 2). The third term helps in minimizing the length of all level sets; the fourth term is the energy-based coupling constraint, used previously, in Eq. 2. The last term enforces the constraint of $|\nabla\phi_i| = 1$, thus helping us avoid explicit redistancing of level sets during the evolution process [10]. Regularized Heaviside and Dirac-delta functions, proposed by Chan and Vese in [1], are also used in our energy functional. $\mu_{in}, \mu_{out}, \nu, \gamma, \eta$ are constants associated with the functional.

$$\begin{aligned}
 E_{fc}(\mathbf{c}_{in}, \mathbf{c}_{out}, \Phi) = & \mu_{in} \left\{ \sum_{i=1}^4 \int_{\Omega} (I - c_{in}^i)^2 H(\phi_i) \right\} d\mathbf{y} + \\
 & \mu_{out} \int_{\Omega} (I - c_{out})^2 \prod_{\substack{i=1 \\ \forall i: H(\phi_i) < 0}}^4 (1 - H(\phi_i)) d\mathbf{y} + \nu \left\{ \sum_{i=1}^4 \int_{\Omega} |\nabla H(\phi_i)| d\mathbf{y} \right\} + \\
 & \gamma \sum_{i=1}^4 \sum_{j=i+1}^4 \int_{\Omega} H(\phi_i) H(\phi_j) d\mathbf{y} + \eta \left\{ \sum_{i=1}^4 \int_{\Omega} \frac{1}{2} (|\nabla\phi_i| - 1)^2 d\mathbf{y} \right\} \quad (2)
 \end{aligned}$$

The four Euler-Lagrange evolution equations associated with the minimization of Eq. 2 are as follows ($i = 1, 2, 3, 4$):

$$\begin{aligned}
 \frac{\partial\phi_i}{\partial t} = & \delta(\phi_i) \left\{ \mu_{in} (I - c_{in}^i)^2 - \mu_{out} (I - c_{out})^2 \prod_{\substack{j=1 \\ \forall j: H(\phi_j) < 0, j \neq i}}^4 (1 - H(\phi_j)) \right. \\
 & \left. - \nu \operatorname{div} \left(\frac{\nabla\phi_i}{|\nabla\phi_i|} \right) + \gamma \sum_{j=1; j \neq i}^4 H(\phi_j) \right\} + \eta \left\{ \Delta\phi_i - \operatorname{div} \left(\frac{\nabla\phi_i}{|\nabla\phi_i|} \right) \right\} \quad (3)
 \end{aligned}$$

where, Δ is the Laplacian operator.

In addition to the energy-based coupling technique of ZZNLS [3] to penalize overlaps between level sets, we use an explicit topological coupling technique. First, we compute $\delta(\phi_i); i \in [1, 4]$. As we use a narrow-band approach (i.e., $\delta(\phi_i) > s_{thresh}$) to update the level set curves we check the saliency of $\delta(\phi_i)$ i.e., $\delta(\phi_i) > \delta(\phi_j); j \neq i$. This helps us identify pixels on the front of the current level set that may lie on narrow-band fronts of other level sets. A pixel on the front of a current level set is updated only if this saliency test is satisfied. If however a “collision” is detected between cells, then the evolution of level sets near the “collision” region stops. To speed up convergence as in [11] and [12] we use level set segmentation from a previous frame as an initial estimate for a current frame.

For details on implementing the four-level set algorithm, we direct the reader to [6].

3 Multiple Cell Tracking Using Correspondence Graphs

Tracking is a fundamental step in the analysis of long term behavior of migrating cells. In this section we present a detection based cell tracking algorithm that extends our previous work in [13]. Tracking is done by resolving frame to frame correspondences between multiple cells segmented using active contour level set methods as described in Sec.2.

3.1 Summary of Tracking Algorithm

1. For each frame $I(\mathbf{y}, t)$ at time t , the tracking module receives four foreground mask layers, $\Omega_k(t)$, that correspond to level sets from the “four-color” segmentation algorithm. In order to refine cell boundaries and remove spurious regions, morphological operations (e.g., opening, closing) *may* be performed on these foreground layers.
2. Connected component analysis is performed on all refined foreground layers such that each $\Omega_k(t)$ is partitioned into n_k disjoint regions

$$\Omega_k(t) = \{\Omega_{k,1}(t), \Omega_{k,2}(t), \dots, \Omega_{k,n_k}(t)\}$$

that ideally correspond to n_k individual cells.

3. For each disjoint region $\Omega_{k,i}$, features such as bounding box, centroid, area, and support map are extracted. Region information from all four layers are combined and relabeled as

$$\Omega_i(t), i \in [1 \dots n], \text{ and } n = \sum_{k=1}^4 n_k$$

without merging connected regions from different foreground layers. Keeping regions from different foreground layers distinct, even when they are spatially connected, preserves the identities of previously disjoint cells, thus preventing false trajectory merges.

4. Relabeled region information is arranged in an “Object-Match” graph structure \mathcal{O}_{GR} that is used in tracking (Fig. ??). Nodes in the graph represent objects $\Omega_i(t)$, while edges represent object correspondences.
5. Correspondence analysis searches for potential object (cell) matches in consecutive frames. \mathcal{O}_{GR} is updated by linking nodes corresponding to objects in frame $I(\mathbf{y}, t)$ with nodes of potential corresponding objects in frame $I(\mathbf{y}, t - 1)$. The confidence value $\mathcal{C}_M(i, j)$ for each match is stored with each link. This is explained further in the following sub-section.
6. The trajectory generation and validation module analyses the \mathcal{O}_{GR} graph and generates valid cell trajectories.

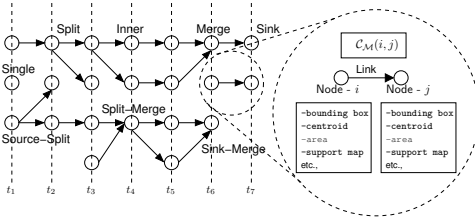


Fig. 1. An Object-Match graph structure, OGR used in cell tracking. Nodes represent detected cells and associated cell features, while links represent frame-to-frame cell correspondences and associated match confidence values.

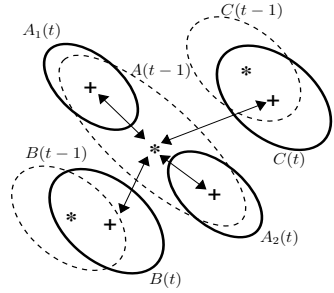


Fig. 2. Centroid distances during cell division. Cell A at time $t - 1$ divides into cells A_1 and A_2 at time t . Centroid distances from the parent A to its children $\overline{AA_1}$ and $\overline{AA_2}$ are comparable in magnitude to centroid distances to its neighbors \overline{AB} and \overline{AC} resulting in ambiguities.

3.2 Cell-to-Cell Correspondence

Cell-to-cell matching (correspondence) is performed using a multi-stage overlap distance \mathcal{D}_{MOD} , which consists of three distinct distance functions $\mathcal{D}_{BB\mathcal{X}}$, \mathcal{D}_{MSK} , and \mathcal{D}_{OLP} for three different ranges of cell motion.

- The *inter-bounding-box distance* $\mathcal{D}_{BB\mathcal{X}}$ quantifies the long-range displacement between two regions (cells) in consecutive frames, by the distance between their bounding boxes defined as the minimum Euclidean distance between corner pairs of the two bounding boxes.

Let $\Lambda[\mathcal{BB}\mathcal{X}(\Omega_i(t)), k]$ and $\Lambda[\mathcal{BB}\mathcal{X}(\Omega_j(t-1)), l]$, indicate the k^{th} and l^{th} corners of the bounding box covering $\Omega_i(t)$ and $\Omega_j(t-1)$. $\mathcal{D}_{BB\mathcal{X}}$ can then be defined as

$$\mathcal{D}_{BB\mathcal{X}}(\Omega_i, \Omega_j) = \eta_b \min_{k,l} \left\{ \left\| \Lambda[\mathcal{BB}\mathcal{X}(\Omega_i), k] - \Lambda[\mathcal{BB}\mathcal{X}(\Omega_j), l] \right\| \right\} \quad (4)$$

where, η_b is a constant.

- The *inter-mask distance* \mathcal{D}_{MSK} quantifies the mid-range displacement between two regions (cells) in consecutive frames, by the minimum contour to contour distance. This distance is computed in terms of the minimum number of dilations needed to overlap the two regions as

$$\mathcal{D}_{MSK}(\Omega_i, \Omega_j) = \eta_m \arg \min_k \left\{ (\Omega_i \oplus_k s_I) \cap \Omega_j \neq \emptyset \right\} \quad (5)$$

where, \oplus_k denotes k -times dilation, s_I denotes unit structuring element, and η_m is a constant.

- The *tonal-weighted overlap distance* $\mathcal{D}_{\mathcal{OLP}}$ quantifies the small-range displacement between two regions (cells) by the degree of their overlap, in terms of shape and tonal dissimilarities. In order to emphasize overlap in nuclei (i.e., dark regions with low intensity values), and to de-emphasize cytoplasm overlap (i.e., light regions with high intensity values), overlapping and non-overlapping regions are weighted by local tonal differences between two regions as

$$\mathcal{D}_{\mathcal{OLP}}(\Omega_i, \Omega_j) = \eta_o \left\{ \int_{\Omega_i \setminus \Omega_j} (1 - I_i(\mathbf{y})) d\mathbf{y} + \int_{\Omega_j \setminus \Omega_i} (1 - I_j(\mathbf{y})) d\mathbf{y} + \int_{\Omega_i \cap \Omega_j} |I_i(\mathbf{y}) - I_j(\mathbf{y})| d\mathbf{y} \right\} / \left\{ \int_{\Omega_i} I_i(\mathbf{y}) d\mathbf{y} + \int_{\Omega_j} I_j(\mathbf{y}) d\mathbf{y} \right\} \quad (6)$$

where, the intensity images $I_i(\mathbf{y}) = I_i(\mathbf{y}, t)$ and $I_j(\mathbf{y}) = I_j(\mathbf{y}, t - 1)$, and are scaled such that $I \in [0, 1]$. η_o is a constant. The first two terms in the numerator of Eq. 6 account for the distance due to uncovered regions in frames at time instants t and $t - 1$, respectively. The complement of intensity images are used to obtain higher distances for uncovered low intensity regions (i.e., nuclei). The third term in the numerator accounts for the intensity dissimilarity within the overlapping region. The denominator is used to normalize the distance by the area of the two cells being compared.

$\mathcal{D}_{\mathcal{MOD}}$ can be assigned any of the three distance measures described above, as per the following rules,

$$\mathcal{D}_{\mathcal{MOD}}(\Omega_i, \Omega_j) = \begin{cases} \mathcal{D}_{\mathcal{BBX}} & \text{if } \mathcal{BBX}(\Omega_i) \cap \mathcal{BBX}(\Omega_j) = \emptyset \\ \mathcal{D}_{\mathcal{MSK}} & \text{if } \Omega_i \cap \Omega_j = \emptyset \\ \mathcal{D}_{\mathcal{OLP}} & \text{otherwise} \end{cases} \quad (7)$$

The proposed multi-stage distance measure depends on size and shape similarity of the compared regions, besides their proximity, and thus have several advantages over the widely used centroid distance measure.

A particularly important case for cell tracking is *mitosis* (i.e., cell division). During mitosis, epithelial cells often become elongated, subsequently splitting across the minor axis. This produces a big increase in the centroid distance and the distances between a cell and its children become comparable to the distances between a cell and its neighboring cells (Fig. 2). In such a scenario, a low gating threshold would result in parent-to-children matches being discarded, resulting in discontinuities in cell trajectories. However, if a high gating threshold is used, correspondence ambiguities may arise. The proposed multi-stage overlap distance measure overcomes these problems.

In addition to object separation based measures, shape (contour) similarity metrics can also be added to the distances described above as additional matching criteria. Since for cell tracking we are primarily interested in the displacement parameter, those metrics are not used in this study but they will be considered in future.

During tracking a match matrix $\overline{\overline{\mathcal{M}}}$ and a confidence matrix $\overline{\overline{\mathcal{C}_{\mathcal{M}}}}$ for each frame, $I(\mathbf{y}, t)$, are produced. $\mathcal{M}(\Omega_i, \Omega_j)$ indicates whether the i^{th} object in $I(\mathbf{y}, t)$, corresponds to the j^{th} object in $I(\mathbf{y}, t - 1)$. $\mathcal{C}_{\mathcal{M}}(\Omega_i, \Omega_j)$ indicates the confidence of this match and consist of two components,

- *Similarity confidence*, $\mathcal{C}_{SIM}(\Omega_i, \Omega_j)$, is a measure of the similarity between the matched objects and is defined as

$$\mathcal{C}_{SIM}(\Omega_i, \Omega_j) = 1 - \frac{\mathcal{D}_{MOD}(\Omega_i, \Omega_j)}{\mathcal{D}_{MOD}^{\max}} \quad (8)$$

where, \mathcal{D}_{MOD}^{\max} is a user defined constant used to normalize \mathcal{D}_{MOD} .

- *Separation confidence*, $\mathcal{C}_{SEP}(\Omega_i, \Omega_j)$, measures the competition between possible matches for the current object and is defined as

$$\mathcal{C}_{SEP}(\Omega_i, \Omega_j) = \begin{cases} 1 & \text{no competitor,} \\ 0.5 \left\{ 1 - \frac{\left(\mathcal{D}_{MOD}(\Omega_i, \Omega_j) - \mathcal{D}_{MOD}(\Omega_i, \Omega_j^*) \right)}{\max\left(\mathcal{D}_{MOD}(\Omega_i, \Omega_j), \mathcal{D}_{MOD}(\Omega_i, \Omega_j^*) \right)} \right\} & \text{otherwise} \end{cases} \quad (9)$$

where, Ω_j indicates the current candidate being compared with Ω_i , and Ω_j^* is its closest competitor in terms of distance. This measure favors matches without competitors, and matches with competitors having higher distances.

Unfeasible correspondences are eliminated using confidence values. Absolute pruning eliminates matches whose confidence values are below a certain threshold, while relative pruning eliminates matches whose confidence values are below a percentage of the confidence for the best match.

3.3 Trajectory Generation and Validation

Trajectory segments are generated from \mathcal{O}_{GR} using a multi-hypothesis testing approach with delayed decision. Besides one-to-one object matches, the proposed tracking algorithm supports many-to-one, one-to-many, many-to-many, one-to-none, or none-to-one matches that may result from false detections or associations, segmentation errors, occlusion, entering, exiting, or division of cells.

The segment generation module analyzes match information by classifying the nodes of \mathcal{O}_{GR} (cells) into nine types; single, source, source-split, sink, inner, split, sink-merge, merge-split, merge, based on the number of parent and child objects.

A data structure (*Segment-List*) is formed by identifying and organizing a linked list (*Trajectory-Segments*) of inner objects starting with a source or split type cell and ending with a merge or sink type cell. Extracted segments are labeled using a method similar to connected component labeling.

Not all the detected segments correspond to actual cell trajectories. Trajectory validation unit checks the validity of each segment based on criteria such as duration, length, linearity, size of the corresponding object, parent and children segments etc. and filters out invalid segments. Trajectories are formed by linking unfiltered segments sharing the same label. Discontinuity resolution is also done in this unit using Kalman filter prediction.

4 Results and Analysis

The proposed cell segmentation and tracking system has been tested on a wound healing image sequence consisting of 136 frames of dimensions 300×300 ($40 \mu\text{m} \times 40 \mu\text{m}$)

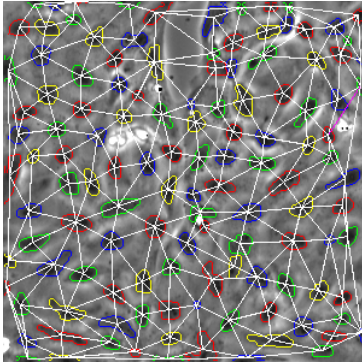


Fig. 3. Segmentation mask obtained using our four-level set formulation with an explicit topological coupling constraint and the associated Delaunay graph for frame #136

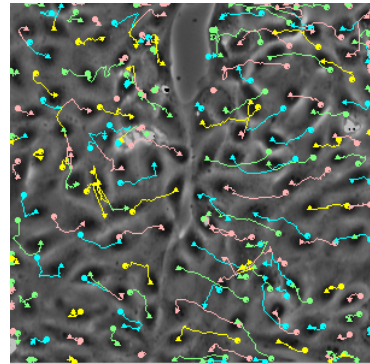


Fig. 4. Trajectories obtained using the described tracking algorithm on masks from our four-level set formulation with an explicit topological coupling constraint

Table 1. Tracking results, when using three different segmentation algorithms. The number of frames in the sequence = 136, with dimensions of each frame equal to 300×300 .

	R-P	T-O	T-S	T-M	T-A	T-D
CV2LS	85%	16748	44	16	14	23
NBP4LS-EC	85%	16732	33	2	29	33
NBP4LS-ETC	85%	17161	22	0	25	20

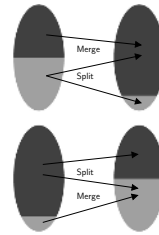


Fig. 5. False splits and merges that result from fragmentation in cells

with image intensities $I \in [0, 255]$. The sequence has been obtained using a monolayer of cultured pig epithelial cells, as described by Salaycik *et al.*, in [14]. Images were sampled uniformly over a 9:00:48 hour period and acquired using a phase contrast microscope, with a $10\times$ objective lens, and at a resolution of approximately $0.13\mu\text{m}$ per pixel.

Three segmentation algorithms have been implemented: a multi-phase Chan and Vese level set algorithm (CV2LS); our four-level set algorithm with only energy-based coupling (NBP4LS-EC); and our four-level set algorithm with energy-based and explicit topological coupling NBP4LS-ETC. The tracking algorithm described in Sec. 3 has been applied to the three sets of masks obtained from these segmentation algorithms, and the results have been compared. For all three segmentation algorithms the following parameters have been used: $\mu_{in} = 1, \mu_{out} = 1, \nu = 1.0/(255.0)^2$ and the number of iterations for each frame has been set to a fixed number $K = 15$. For the

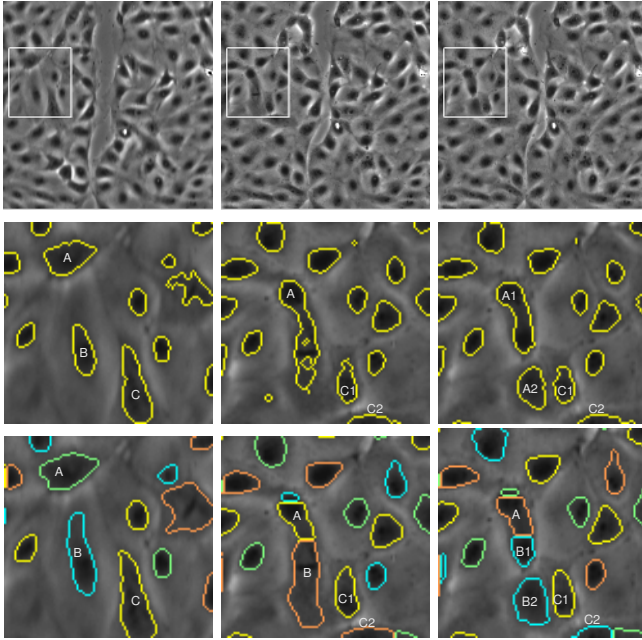


Fig. 6. Evolution of three cells labelled as A,B, and C. First row: Original frames #78,#103, and #111. Second and third rows: segmentation results for CV2LS and NBP4LS-ETC algorithms respectively. Segmentation results are zoomed to the region of interest ([70,10]-[170,110]) marked on the original frames. In row 3, contour colors associated with each object are changing from frame to frame due to re-coloring of the neighborhood graph in order to reflect changes in the neighborhood relationships.

segmentation algorithms with energy-based coupling constraint γ has been set to 0.1. During tracking, a relative pruning rate of 85% has been used and matches with confidence values below 85% of \mathcal{C}_{BST} have been pruned. (\mathcal{C}_{BST} indicates the confidence of the best match for the current object). The tracking results have been filtered by object size and segment length. Size threshold T_{os} has been set to 30 pixels, and duration threshold T_{sl} has been set to 5 frames. Incomplete objects at the image borders (i.e., within 20 pixels from each side) have been excluded from the statistics.

Representative results for our segmentation (NBP4LS-ETC) and tracking algorithms are given in Figures 3 and 4. Figure 3 shows segmented cells and their neighborhood relationships in the form of a Delaunay graph. Figure 4 shows the cell trajectories obtained after tracking. Table 1 shows tracking results obtained using the tracking algorithm described in Section 3, on masks produced by three different segmentation algorithms, CV2LS, NBP4LS-EC and NBP4LS-ETC. **T-O** indicates the total number of disjoint objects (i.e., cells) detected in all frames of the image sequence. **T-S** indicates the number of trajectory segments that split from another trajectory segment. Such splits may result from cell division, or fragmentation during segmentation. **T-M** indicates number of trajectory segments that merge with other trajectory segments. **T-A** and **T-D** indicate numbers of trajectory segments that appear or disappear, unexpect-

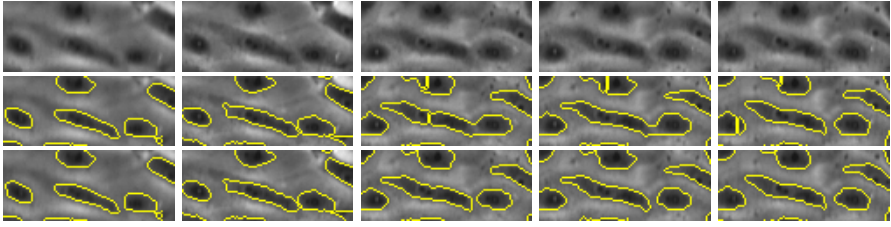


Fig. 7. Segmentation results for the frames #72, #79, #110, #113 and #118, zoomed to the region of interest ([256,13]-[295,110]). First row: original sequence. Second row: segmentation using NBP4LS-EC algorithm. Third row: segmentation using NBP4LS-ETC algorithm (contours shown without level set color).

edly. These may occur when no evidence of cells in the current frame exists in previous frames, or when no match exists in future frames for cells present in the current frame. The NBP4LS-ETC algorithm does a better job than either the NBP4LS-EC or the CV2LS algorithm in preventing false-merges of cells. Fragmentation is a problem with both the NBP4LS-ETC and NBP4LS-EC algorithms. But as shown in Table 1, the NBP4LS-ETC results in a smaller number of splits than either NBP4LS-EC or CV2LS algorithms. As shown in Fig. 5, false splits and merges may arise when a cell is fragmented in the current frame, or becomes fragmented in a future frame. Proposed method removes fragments at the image level through post-processing of small objects using morphology and at the trajectory level through filtering short segments.

Figures 6 and 7 show two cases that demonstrate the advantage of our NBP4LS-ETC method. Figure 6 shows the evolution of three cells labeled as A,B, and C. In the sequence from frame #78 to #111 two of the cells undergo mitosis (cell division). Cell C divides first followed by B; between frames #78 and #103 cell C splits into children cells C1 and C2; between frames #103 and #111 cell B splits into children cells B1 and B2. Both algorithms CV2LS (2nd row), and NBP4LS-ETC (3rd row) correctly identify the mitosis event of cell C. But only NBP4LS-ETC correctly identifies the mitosis event of cell B. Using the CV2LS method, at frame #103 cell B that is clearly distinct in frame #78, gets falsely merged with cell A, because cell B becomes indistinct (the nucleus is not as thick as preparation for DNA replication) and at the same time the region between cells A and B gets darker as these cells move closer together. This false merge causes additional tracking complications when cell B undergoes subsequent mitosis. As seen in frame #111, cell A appears to undergo mitosis and is incorrectly associated with the children A1 and A2. Cell A2 is associated with parent cell A instead of parent B due to a correspondence error; it should be labeled as cell B2. Object A1 is associated with parent object A due to both correspondence and segmentation errors; the segmentation of object A1 merges two actual cells (labeled A and B1 in row 3) which leads to the association error during tracking. However using the NBP4LS-ETC algorithm the explicit topological coupling constraint prevents cells A and B from merging and the mitosis of cell B is correctly detected and tracked. Effects of fragmentation such as the small region above cell A in frames #103 and #111 are handled by postprocessing and trajectory filtering and did not cause any tracking errors.

Figure 7 depicts an “absorption event” where one level set pushes a neighboring

level set out of its own path of evolution. Using only an energy-based coupling term (2^{nd} row - NBP4LS-EC) leads to the shifting of the boundary between adjacent objects. Ultimately one object “absorbs” the other without a merge event (where adjacent level sets merge together during contour evolution). Using the proposed additional explicit topological coupling term as described, the location of the boundary is maintained which prevents absorption. The absorption process may be order dependent but we have found that randomization of level set processing order is not sufficient to prevent absorption.

5 Conclusion

We have presented an automated cell segmentation and tracking system designed to study migration of cells imaged with a phase contrast microscope. Cells that move into the field-of-view, arise from cell division or disappear due to apoptosis are reliably segmented and tracked by this system. The novel four-color level set formulation introduced to deal with the false-merge problem in segmentation and tracking of dense cell clusters, is very scalable and significantly reduces the computational complexity of N -level set formulation of Zhang *et al.*, [3]. Experimental results show that segmentation with the proposed four-level set formulation, with an explicit topological coupling constraint, greatly improves accuracy of trajectories obtained during cell tracking. Further research on trajectory validation and behavior analysis is currently in progress.

References

1. T.Chan, L.Vese: Active contours without edges. *IEEE Trans. Image Process.* **10** (2001) 266–277
2. L.Vese, T.Chan: A multiphase level set framework for image segmentation using the Mumford and Shah model. *Intern. J. Comput. Vis.* **50** (2002) 271–293
3. B.Zhang, C.Zimmer, J.-C.Olivo-Marin: Tracking fluorescent cells with coupled geometric active contours. In: *Proc. 2nd IEEE Int. Symp. Biomed. Imaging (ISBI)*, Arlington, VA (2004) 476–479
4. A.Dufour, V.Shinin, S.Tajbakhsh, N.Guillén-Aghion, J.-C.Olivo-Marin, C.Zimmer: Segmenting and tracking fluorescent cells in dynamic 3-D microscopy with coupled active surfaces. *IEEE Trans. Image Process.* **14** (2005) 1396–1410
5. C.Zimmer, J.-C.Olivo-Marin: Coupled parametric active contours. *IEEE Trans. Pattern Anal. Machine Intell.* **27** (2005) 1838–1842
6. S.Nath, K.Palaniappan, F.Bunyak: Cell segmentation using coupled level sets and graph-vertex coloring. In R.Larsen and M.Nielsen and J.Sporring, ed.: *LNCS - Proc. MICCAI 2006*. Springer-Verlag (2006)
7. K.Appel, W.Haken: Every planar map is four colorable. Part I. discharging. *Illinois. J. Math.* **21** (1977) 429–490
8. K.Appel, W.Haken, J.Koch: Every planar map is four colorable. Part II. reducibility. *Illinois. J. Math.* **21** (1977) 491–567
9. N.Robertson, D.P.Sanders, P.D.Seymour, R.Thomas: The four color theorem. *J. Combin. Theory, Ser. B* **70** (1997) 2–44
10. C.Li, C.Xu, C.Gui, D.Fox: Level set evolution without re-initialization: A new variational formulation. In: *Proc. IEEE Conf. Computer Vision Pattern Recognition*. Volume 1. (2005) 430–436

11. D.P.Mukherjee, N.Ray, S.T.Acton: Level set analysis for leukocyte detection and tracking. *IEEE Trans. Image Process.* **13** (2001) 562–672
12. N.Paragiosis, R.Deriche: Geodesic active contours and level sets for the detection and tracking of moving objects. *IEEE Trans. Pattern Anal. Machine Intell.* **22** (2000) 266–280
13. F.Bunyak, K.Palaniappan, S.K.Nath, T.I.Baskin, G.Dong: Quantitive cell motility for *in vitro* wound healing using level set-based active contour tracking. In: *Proc. 3rd IEEE Int. Symp. Biomed. Imaging (ISBI)*, Arlington, VA (2006)
14. K.J.Salaycik, C.J.Fagerstrom, K.Murthy, U.S.Tulu, P.Wadsworth: Quantification of micro-tubule nucleation growth and dynamics in wound-edge cells. *J. Cell Sci.* **118** (2005) 4113–4122

Robust Visual Identifier for Cropped Natural Photos

Ik-Hwan Cho¹, A-Young Cho¹, Hae-Kwang Kim²,
Weon-Geun Oh³, and Dong-Seok Jeong¹

¹ Department of Electronic Engineering, Inha University,
Yonghyun-Dong, Nam-Gu, Incheon, Republic of Korea
{teddydino, ayoung}@inhaian.net,
dsjeong@inha.ac.kr

² Department of Software Engineering, Sejong University,
98 Gunja-Dong, Gwangjin-Gu, Seoul, Republic of Korea
hkkim@sejong.ac.kr

³ Electronics and Telecommunications Research Institute,
161 Gajeong-Dong, Yuseong-Gu, Daejeon, Republic of Korea
owg@etri.re.kr

Abstract. The cropping of image is one of most popular functions in current image editing software for general digital camera users. And image-based identifier system is needed for wide distribution of digital image products. In this paper, we propose new concept of visual identifier for digital photos and visual identifier system structure robust against especially cropped natural photos. Visual identifier is new concept with different ground truth rather than retrieval. In the proposed identifier system, local corner is used as main co-location position and local gradient histogram is applied to describe feature in each position. And for robust matching we use simple random sample consensus method. Since image cropping can be considered as a kind of translation, linear model is sufficient as geometric transform model. For experiment we make 11 kinds of modifications of original images and evaluate performance of the proposed algorithm. From experiment results, our proposed algorithm shows better performance relative to previous MPEG-7 visual descriptors.

1 Introduction

As more and more multimedia contents are available, efficient and effective management of multimedia contents are required. As the result of recent many years of work, MPEG-7 visual standard now has many visual descriptors on the basis of color, texture, shape and motion [1].

A visual data can be modified by editing, transcoding, etc. through its life publishing many versions with or without intention for legal or illegal purposes. There is wide usage of visual identifier descriptor which will identify these versions of one visual data from other visual data.

One important application is for illegal usage tracking. The rights owner of a photo wants to track where at the Internet his photo copies are. One possible scenario is that an automatic illegal usage tracking agent will gather visual data from the Internet and

extract each visual identifier for a visual content and compare it with already extracted its visual identifier descriptor at its hand.

Current visual descriptors based on MPEG-7 elements are may be used for a visual identifier descriptor, however, the test conditions for the conventional MPEG-7 visual descriptors are set for only image and video retrieval, not for identifier. For the ground truth, similar visual data are used even they are not from the same original video data.

In this paper, we propose visual identifier system robust against especially photo cropping modification. Cropping of image or natural photo is one of the most popular and frequent functions used in general imaging software like Adobe Photoshop [2]. Therefore cropped image is created easily by not only its owner, but also the others. Hence cropped version of original photo can be first candidate of various modification possibilities.

This paper is organized as following; section 2 describes visual identifier and hit ratio for evaluation of performance. And section 3 explains feature extraction and section 4 explains similarity matching process. In section 5 experiment results are depicted and section 6 leads discussion for this paper and finally section 7 concludes.

2 Visual Identifier and Hit Ratio

In this paper, visual identifier is extended concept of visual retrieval. So the proposed concept of visual identifier can be described from conventional visual retrieval [3]. Fig. 1 shows the concept of visual identifier.

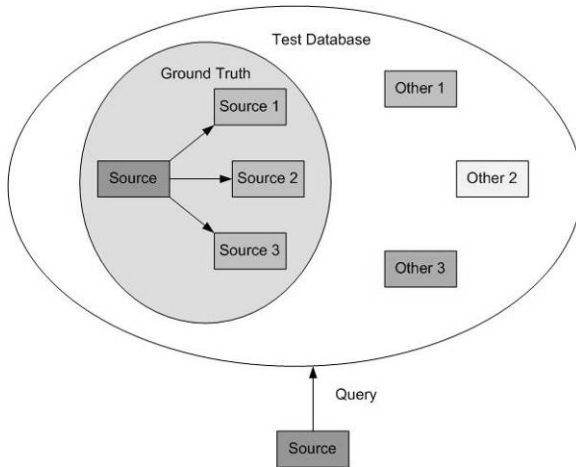


Fig. 1. Test database and ground truth for visual identifier

In general image retrieval system, test database has a lot of image data which is independent ground truth as itself. As some query image is asked to retrieval system ideal system must retrieve images in according to minimum similarity distance of

specific criteria. For this purpose MPEG made international standard and its current output is MPEG-7 [1]. MPEG-7 has several audio and visual descriptors for retrieval system and its descriptor has individual criteria and classification in accordance with feature [1], [4]. For example, as using MPEG-7 dominant color system retrieves images which have dominant color similar to query image even if they are actually not related with query image. If there is modified version of original data in test database pool, it become individual ground truth not related with original one. On the contrary visual identifier system has difference in the composition of ground truth. What 'Identifier' means is to discriminate if there is original or modified versions of query data in test database. So ground truth includes original version as well as its modified ones.

The performance is estimated with average hit ratio [5]. For a query i , only the $(N+1)$ with the highest similarities are counted for the performance evaluation where N is the number of modified versions. The hit number h_i is the number of ground truth contents that are retrieved. The hit ratio R_i for the query is calculated as $R_i = h_i / (N+1)$. The average hit ratio for all the queries are calculated as its performance index.

3 Feature Extraction

In this paper, the proposed visual identifier system consists of two parts, feature extraction and similarity matching. In feature extraction part, co-location position is set up and features are extracted from each image and extracted features are encoded into binary type file. Fig. 2 shows the simple block diagram for feature extraction process. To evaluate performance of the proposed visual identifier, we make test image database which includes original version and several kind of modifications.

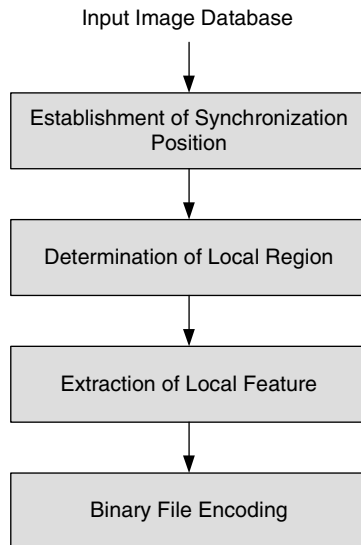


Fig. 2. Overall block diagram of feature extraction process

3.1 Establishment of Co-location Position

In ideal visual identifier system, modified version like cropping must be considered as query or reference image data. For example, cropped image has limited information relative to original version. Consequently it is necessary to co-locate same positions to identify the correlation between original and modified versions. In this paper, interesting points are used to determine co-location position to be described and to obtain corner as interesting point Harris corner point detection method is used [6]. Harris corner points are able to be obtained from corner response equation of Eq. (1).

$$\begin{aligned}
 R &= \det M - k(\text{trace } M)^2 \\
 \det M &= \lambda_1 \lambda_2 \\
 \text{trace } M &= \lambda_1 + \lambda_2 \\
 M &= \sum_{x,y} w(x,y) \begin{bmatrix} I_x^2 & I_x I_y \\ I_x I_y & I_y^2 \end{bmatrix}
 \end{aligned} \tag{1}$$

where R is Harris corner response, I_x and I_y is first-derivative of input image I . And w is Gaussian kernel.

In input image, pixels having corner response value R over threshold are classified into candidate interesting points. And final Harris corner points are detected by selecting spatial local maxima pixels from candidate points. The number of interesting points may be controlled by threshold of corner response value. In generally, trade-off exists between number of interesting points (or data size) and performance.

3.2 Determination of Local Region

After obtaining interesting points as co-location positions, we need to determine the size of local region. Real image descriptor is computed from just this limited local region. Other regions not selected as synch position are excluded from description range. Using this value, rectangular local regions are set. And several regions may be overlapped if distance between points is less than region size.

3.3 Extraction of Local Feature

In this contribution, we use local gradient histogram as real descriptor. In prior to real description, local region should be normalized. Local region is smoothed by using Gaussian smooth function. And then gradient amplitude and phase for every normalized pixel is calculated. Eq. (2) represents the calculation of amplitude and phase of each gradient.

$$m(x, y) = \sqrt{(L(x+1, y) - L(x-1, y))^2 + (L(x, y+1) - L(x, y-1))^2} \quad (2)$$

$$\theta(x, y) = \tan^{-1}((L(x, y+1) - L(x, y-1)) / (L(x+1, y) - L(x-1, y)))$$

where $m(x,y)$ is the amplitude of gradient, $\theta(x,y)$ is phase of gradient and $L(x,y)$ is pixel in (x,y) location of local region.

To compute gradient histogram as descriptor, all angles are divided into 36 bins and each bin represents 10 degree of gradient phase. Therefore number of bins of gradient histogram is 36 and amplitude of each bin is the weighted sum of gradient amplitude in according to its phase. In this process, Gaussian function is used as weighted function. Eq. (3) is gradient histogram equation considering weight on center position of local region.

$$GH(i) = \sum_{x=0, y=0}^S w_g(x, y) p(x, y, i) m(x, y)$$

$$w_g(x, y) = \frac{1}{2\pi\sigma^2} e^{-\frac{(x-cx)^2 + (y-cy)^2}{2\sigma^2}} \quad (3)$$

$$p(x, y, i) = \begin{cases} 1, & \text{if index of gradient phase } \theta(x, y) \text{ in position } (x, y) \text{ is equal to } i \\ 0, & \text{otherwise} \end{cases}$$

where $GH(i)$ is weighted gradient histogram for phase index i , w_g is Gaussian weighting function when center position of local region is (cx, cy) . And $\theta(x,y)$ is gradient phase in (x,y) position.

4 Similarity Matching

The ideal visual identifier must determine whether there are original or modified versions of query in target database. In above feature extraction section, features for all images of test database extracted and stored as one binary format. The proposed visual identifier extracts feature from query image and measure distance between query and each reference in database. From distance measure, visual identifier concludes whether ground truth or query exists in database.

In this paper, additional processes including correspondence matching and Random Sample Consensus (RANSAC) are needed to get matching positions between both descriptors [7]. Firstly, the correspondence matching using only spatial pattern of local region finds candidate matching pairs using simple Euclidean distance between gradient histograms. As a next step, RANSAC removes outlier pairs from candidate matching pairs derived in the previous step by considering relation of geometric transform. Finally, we can obtain similarity between two images using distance measure. Fig. 3 represents overall similarity matching process.

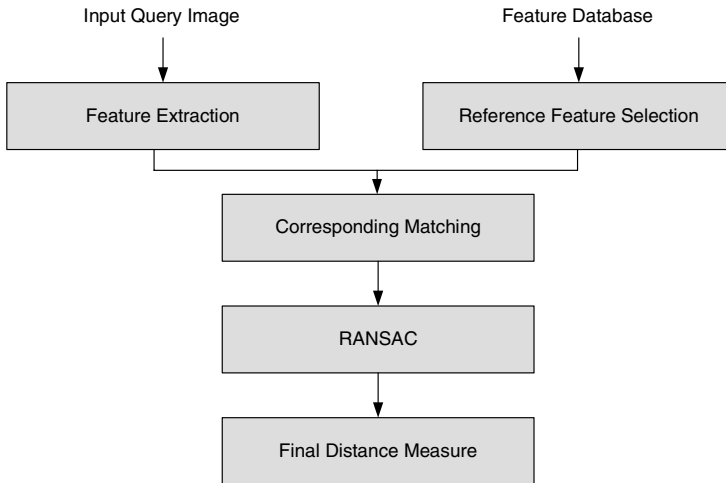


Fig. 3. Block diagram of similarity matching process between query and reference images

4.1 Feature Extraction

For similarity matching, all features is extracted from only query image with same method explained above section and it is not necessary for reference images in database because all features are already extracted and stored as binary format . From feature extraction, co-location positions are obtained and local region size is fixed the same as given in the above section. And in each local region, gradient histogram is extracted as main descriptor.

4.2 Correspondence Matching

For one image, several descriptors are extracted in according to the number of co-location positions and therefore for similarity matching it is necessary to set up matching pairs between descriptor groups in both images. In the proposed method, many local co-location positions utilizing image characteristics are used as corresponding matching positions between both images instead that overall image resolution is not used as matching position for descriptors. Initial matching pairs are obtained by correspondence matching process which is first one of two steps for fixing matching positions. Pseudo code is depicted as following;

```

For (i; all local feature in reference)
  For (j; all local feature in query)
    Dist = Euclidean distance(i,j);
    If (Dist < threshold)
      Add (i,j) to initial matching pairs
    End
  End
End
  
```

As seen in the above pseudo code, initial matching pairs are constructed using individual similarity of local region texture information. It considers neither location

of local regions nor geometric relation. Therefore, in initial matching pairs, we can get not only 1:1 but also 1:n matching pairs.

4.3 RANSAC

RANSAC is the algorithm which predicts the model from the observed data. Throughout RANSAC, we can obtain the relational motion model, especially in image matching. In this paper, we use the correspondence of motion vectors. Pseudo code is represented following;

```

For (i; all motion vectors){
  initialize inlier_index[n]
  For (j; all motion vectors){
    dist =distance (motion vector i- motion vector j);
    If (dist < threshold_dist)
      Ninlier ++;
      inlier_index[j]=1;
  }
  If (Ninlier > Nmv /2 )
    Goto end;
}

//Outlier rejection
For (i; all motion vectors){
  If (inlier_index[i]!=1)
    remove motion vector i;
}

```

In general, inlier is the sample corresponded to model and outlier is the sample not corresponded to model. From RANSAC with initial matching pairs, outlier pairs which may decrease the similarity performance in final matching are removed.

4.4 Final Distance Measure

From final matching pairs, we can calculate the final distance between query and reference. In this contribution, the average Euclidean distance in Eq. (4) is used as the final similarity measure.

$$D = \frac{1}{N_p} \sum_{i=1}^{N_p} |GH_{ref}(i) - GH_{qur}(i)| = \frac{1}{N_p} \sum_{i=1}^{N_p} \sqrt{\sum_{j=1}^{36} (GH_{ref}(i, j) - GH_{qur}(i, j))^2} \quad (4)$$

where $GH_{ref}(i)$ and $GH_{qur}(i)$ are i th gradient histogram vector of reference and query and N_p is total number of matching pairs between reference and query.

5 Experiment Results

The performance of the proposed descriptor and matching process is measured by average hit ratio. Total number of ground truth is 12 since 11 kinds of modification are used. Table 1 shows various modifications used in this contribution.

Table 1. Image modifications for visual identifier test

No.	Modification
O	Original
M ₁	Brightness (+5)
M ₂	Brightness (-5)
M ₃	monochrome
M ₄	JPEG compression (QF:95%)
M ₅	Color reduction (16bit color)
M ₆	Blur 3x3
M ₇	Histogram equalization
M ₈	Crop 90 %
M ₉	Crop 70 %
M ₁₀	Crop 50 %
M ₁₁	Flexible crop

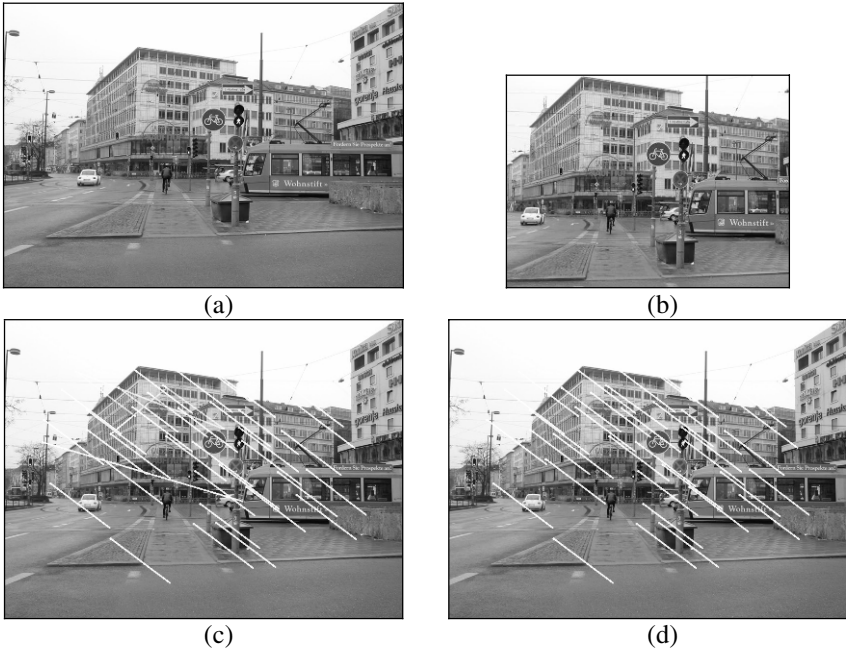


Fig. 4. Matching pairs results between original and cropped images. (a) Original image, (b) cropped image(70%), (c) matching pairs before RANSAC and (d) matching pairs after RANSAC. White line means connection of matched points between reference and query images.

We use 200 original images. Therefore, the total number of test image set including the modified is 3,000. Feature extraction process uses all 3000 images and the matching process uses 750 (50x15=750) as query images. Fig. 4 represents our

experimental results with the performance of edge histogram descriptor. The proposed visual identifier system is implemented by modifying MPEG-7 reference software based on Pentium 2.8GHz Windows XP system.

Simple result of the proposed method is depicted in Fig. 4. Fig. 4(a) is original image and Fig. 4(b) is its cropped image by 70% size of original one. Through feature extraction and similarity matching in both sides, we obtain matching pairs to calculate similarity distance. In Fig. 4(c) which is the output of correspondence matching, wrong matching pairs exist since correspondence matching is based on only local descriptor. After applying RANSAC while considering geometric transform, just correct matching pairs remains and outlier is removed.

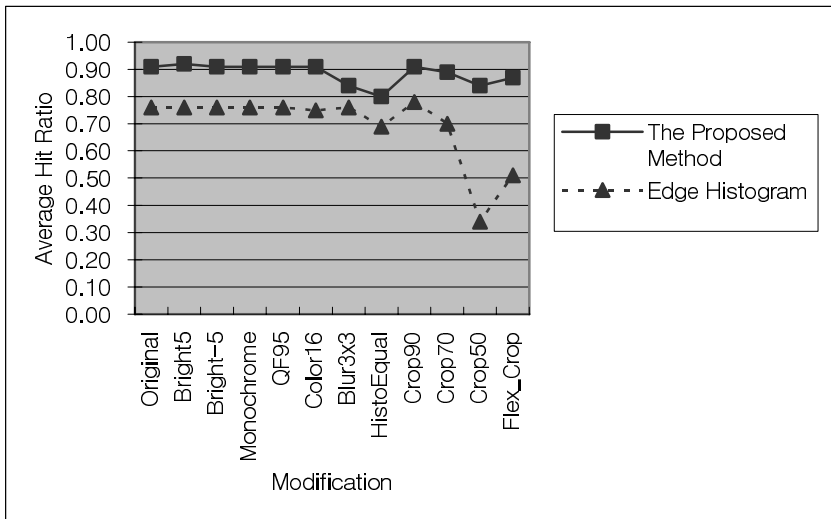


Fig. 5. Average hit ratio of edge histogram descriptor and the proposed method

For generic image processing such as blurring and color changing as seen in Fig. 5, the proposed method and edge histogram descriptor show high performance together and the proposed one is higher. However for cropping operation performance of edge histogram decreases and as cropping rate increases, its decreasing rate is higher.

6 Discussion

In this paper, the visual identifier system robust against especially image cropping is proposed as a new concept. The identifier system consists of two parts which are feature extraction and similarity matching. As a main feature of identifier, local gradient histogram centered in feature point in input image is extracted and local corner points based on Harris corner point detection method are used for feature points. In similarity matching process, Euclidean distance between gradient histograms between reference and query is used as similarity criteria. To complement matching pairs for the measurement of distance, RANSAC based on simple geometric transform like vector is used.

Visual identifier system is needed to be considered differently rather than retrieval since their objects are different each other. In this paper, new concept of visual identifier is proposed and especially it is focused on identifier of image cropping since for recent users cropping of digital images and photos is one of most frequent functions. In cropped image, the amount of information is smaller than original one so that local region based co-location concept is needed to compare characteristics of images. Like edge histogram global region based matching method is not good for cropped image because descriptors between two images are not matched correctly even if its descriptor is powerful and its performance in generic distortion is very good.

As seen Fig. 4, cropped image has limited information and therefore its description is also limited rather than original image. It is different to generic distorted cases like image blurring or color change. In these cases, robust descriptor can extract similar information even if base image is distorted. But, in cases of cropped image, robust descriptor cannot extract any information since base image is removed. From feature information point of view, information of cropped image is subset of one of original image. Therefore co-location between full information and limited one is very important factor in similarity matching.

Main co-location positions exist locally in image and to make matching pairs between several points geometric relation must be concerned. RANSAC is typical method to estimate geometric relation between datasets. In this paper, simple geometric transformation considering only cropping is used for RANSAC because image cropping is one of translation of a part of original data so that geometric transform is just simple translation. If geometric relation is not considered like Fig. 4(c), incorrect matching pairs may exist and final distance is wrong. By the composition of individual components commented above, the proposed visual identifier shows very high performance rather than edge histogram which is best descriptors in conventional MPEG-7 visual descriptors. For cropped images, edge histogram shows very low performance while the proposed method shows high performance for all cases. Especially for cropped by 50% of the original, its resolution is downsized by 1/4 and its image information decrease by 1/4. Even if it has very limited information, the proposed method identifies its ground truth.

7 Conclusions

In this paper, new visual identifier concept and system robust especially cropped image are proposed. For image cropping as frequent image modification, local region-based description method is proposed and correct co-location of features using simple RANSAC leads high performance in identifying ground truth. In conclusion, the proposed method can support the basement of the visual identifier for especially cases of distorted image with limited information.

Acknowledgment

The presented research is supported by Electronics and Telecommunication Research Institute (ETRI).

References

1. ISO/MPEG N4358, Text of ISO/IEC Final Draft International Standard 15938-3 Information Technology - Multimedia Content Description Interface - Part 3 Visual, MPEG Video Group, Sydney, July 2001.
2. <http://www.adobe.com/photoshop/>
3. Weon-Geun Oh, IK-Hwan Cho, A-Young Cho, Hyun-Mi Kim, Dong-Seok Jeong, Hae-Kwang Kim, Sung-Phil Heo(KT), "Feasibility Test of MPEG-7 Visual Descriptors as a Visual Identifier Descriptor", MPEG Doc. No. M12202, Pozan, July 2005.
4. ISO/MPEG N4224, Text of ISO/IEC Final Draft International Standard 15938-4 Information Technology - Multimedia Content Description Interface - Part 4 Audio, MPEG Audio Group, Sydney, July 2001.
5. Jae-Gwi Choi, Weon-Geun Oh, A-Young Cho, Ik-Hwan Cho, Hyun-Mi Kim, Dong-Seok Jeong, Hae-Kwang Kim, " Proposed test conditions for MPEG-7 Visual Core Experiments 6", MPEG Doc. No. M12841, Bangkok, January 2006.
6. C. Harris and M. Stephens, "A combined corner and edge detector", Proc. Alvey Vision Conf., Univ. Manchester, pp. 147-151, 1988.
7. M. A. Fischler and R. C. Bolles, "Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography," Comm. ACM, 24-6 (1981), 381.395.

Affine Epipolar Direction from Two Views of a Planar Contour^{*}

Maria Alberich-Carramiñana¹, Guillem Alenyà², Juan Andrade-Cetto³,
Elisa Martínez⁴, and Carme Torras²

¹ Departament de Matemàtica Aplicada I,
UPC Avda. Diagonal 647, 08028 Barcelona
maria.alberich@upc.es

² Institut de Robòtica i Informàtica Industrial,
CSIC-UPC Llorens i Artigas 4-6, 08028 Barcelona
{galenya, torras}@iri.upc.edu

³ Centre de Visió per Computador, UAB Edifici O,
Campus UAB, 08193 Bellaterra, Spain
cetto@cvc.uab.es

⁴ CiTS La Salle, Universitat Ramon Llull Pge. Bonanova 8, 08022 Barcelona
elisa@salleurl.edu

Abstract. Most approaches to camera motion estimation from image sequences require matching the projections of at least 4 non-coplanar points in the scene. The case of points lying on a plane has only recently been addressed, using mainly projective cameras. We here study what can be recovered from two uncalibrated views of a *planar contour* under *affine* viewing conditions. We prove that the affine epipolar direction can be recovered provided camera motion is free of cyclorotation. The proposed method consists of two steps: 1) computing the affinity between two views by tracking a planar contour, and 2) recovering the epipolar direction by solving a second-order equation on the affinity parameters. Two sets of experiments were performed to evaluate the accuracy of the method. First, synthetic image streams were used to assess the sensitivity of the method to controlled changes in viewing conditions and to image noise. Then, the method was tested under more realistic conditions by using a robot arm to obtain calibrated image streams, which permit comparing our results to ground truth.

1 Introduction

Recovering camera motion from image streams is an important task in a range of applications including robot navigation and manipulation. This requires a measure of the visual motion on the image plane and a model that relates this motion to the real 3D motion. Most of the existing work on motion recovery relies on a set of point matches to measure visual motion, and, depending on the acquisition conditions, different camera models have been used to emulate the imaging process [1,2]. The full perspective model (the pinhole camera), in

^{*} This work is partially funded by the EU PACO-PLUS project FP6-2004-IST-4-27657.

either its calibrated (perspective camera) or uncalibrated (projective camera) versions, has proved to be too general when perspective effects diminish. Under weak-perspective viewing conditions (small field of view, or small depth variation in the scene along the line of sight compared to its average distance from the camera), simplified camera models, such as orthographic, scaled-orthographic or their generalization for the uncalibrated case, the affine camera model, provide an advantageous approximation to the pinhole camera, which avoids computing ill-conditioned parameters by explicitly incorporating the ambiguities due to weak perspective into the model.

This paper addresses the motion estimation problem in the context of an affine camera using active contours to measure visual motion. There are several previous motion estimation methods based on affine cameras [3,4]. A common feature of these algorithms is that they require the matching of at least four non-coplanar points and fail for planar structures [5]. The particular case of features lying on planes has not been analyzed in detail thus far. The formulation of this problem is the core of the present paper.

It is well known that two views of a plane are related by a collineation under full perspective projection. Several authors have used this fact to propose algorithms for camera calibration [6], self-calibration [7,8], or extraction of structure and motion from uncalibrated views of points on planes [9] or of planar curves [10]. However, when perspective effects diminish, the relationship between two views of a planar structure becomes an affinity, which invalidates the methods based on collineations.

Following the stratified analysis of motion for affine viewing conditions introduced by Koenderink and van Doorn [3] and revisited by Shapiro et al. [4], we first explore what information of the affine epipolar geometry can be inferred from the affine deformation of the projection of a rigid and planar contour in two weak-perspective views. This sets the basis to derive the motion parameters in a second stage. We show that, under a 3D motion free of cyclorotation, the epipolar direction can be recovered by relating the two affine views of the contour. A series of experiments is performed to test the sensitivity of the method to the different conditions imposed.

The paper is organized as follows. Section 2 contains the analytic study of two weak-perspective views and provides the basis for the recovery of the epipolar direction. Section 3 explains how the parameters of the affinity relating the two views are extracted in our implementation, based on a contour tracker. Section 4 is devoted to experimentation, using both synthetic and real image streams. Finally, Section 5 summarizes our contribution and gives some prospects for future work.

2 Analytic Study of Two Weak-Perspective Views

2.1 The Camera Model

We assume that the scene object is stationary and that the camera translates by \mathbf{T} and rotates by \mathbf{R} around the object, and possibly zooms. A new affine

coordinate frame associated with a second camera is given by the rows of \mathbf{R} and the new origin lies at $-\mathbf{R}^T \mathbf{T}$ thus a point in this second camera is given by the expression

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \frac{f'}{Z'_{\text{ave}}} \begin{bmatrix} X' \\ Y' \end{bmatrix}, \tag{1}$$

where $[X, Y, Z]^T = \mathbf{R}[X', Y', Z']^T + \mathbf{T}$, f' is the new focal length, and Z'_{ave} is the average distance to the object from the second camera.

Consider the equation $aX + bY + c = Z$ of a world plane \mathcal{S} . Then the two views of the coplanar scene are related by the affinity given by

$$\begin{bmatrix} x' \\ y' \end{bmatrix} = \mathbf{M} \begin{bmatrix} x \\ y \end{bmatrix} + \mathbf{t}, \tag{2}$$

with

$$\mathbf{M} = s \frac{f'}{f} \begin{bmatrix} R_{1,1} + aR_{1,3} & R_{1,2} + bR_{1,3} \\ R_{2,1} + aR_{2,3} & R_{2,2} + bR_{2,3} \end{bmatrix}, \tag{3}$$

$$\mathbf{t} = -\frac{f'}{Z'_{\text{ave}}} \begin{bmatrix} R_{1,1} & R_{1,2} & R_{1,3} \\ R_{2,1} & R_{2,2} & R_{2,3} \end{bmatrix} \begin{bmatrix} T_x \\ T_y \\ T_z \end{bmatrix} + c \begin{bmatrix} R_{1,3} \\ R_{2,3} \end{bmatrix}, \tag{4}$$

and where $s = Z_{\text{ave}}/Z'_{\text{ave}}$ is the scale factor that accounts for depth variation ($s > 1$ if the second camera approaches the scene object, and $s < 1$ if it departs from it), and $R_{i,j}$ are the elements of the rotation matrix \mathbf{R} .

A direction $\mathbf{v} = [x, y]^T$ of the first image \mathcal{R} is mapped by the above affinity to the direction $\mathbf{M}\mathbf{v}$ of the second image \mathcal{R}' . Since the affine references chosen in the two cameras match by the displacement, we can superpose the two images and it has sense to consider directions invariant by \mathbf{M} .

2.2 Recovery of the Epipolar Direction

Consider an orthonormal coordinate frame associated to the first image (for instance, normalized pixel coordinates, when aspect ratio and skew are known). The rotation matrix about the unit axis $[\cos \alpha, \sin \alpha, 0]^T$ and angle ρ has the form

$$\mathbf{R} = \begin{bmatrix} (1 - \cos \rho) \cos^2 \alpha + \cos \rho & \cos \alpha \sin \alpha (1 - \cos \rho) & \sin \alpha \sin \rho \\ \cos \alpha \sin \alpha (1 - \cos \rho) & (1 - \cos \rho) \sin^2 \alpha + \cos \rho - \cos \alpha \sin \rho \\ -\sin \alpha \sin \rho & \cos \alpha \sin \rho & \cos \rho \end{bmatrix}. \tag{5}$$

Hence, the matrix \mathbf{M} is

$$\mathbf{M} = s \frac{f'}{f} \begin{bmatrix} (1 - \cos \rho) \cos^2 \alpha & \cos \alpha \sin \alpha (1 - \cos \rho) \\ + \cos \rho + a \sin \alpha \sin \rho & + b \sin \alpha \sin \rho \\ \cos \alpha \sin \alpha (1 - \cos \rho) & (1 - \cos \rho) \sin^2 \alpha \\ -a \cos \alpha \sin \rho & + \cos \rho - b \cos \alpha \sin \rho \end{bmatrix}, \tag{6}$$

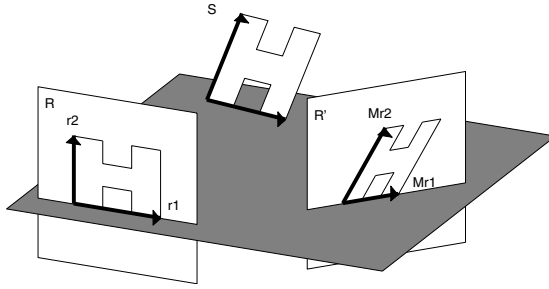


Fig. 1. Graphic illustration of Theorem 1. See text for details.

where $\mathbf{a} = [\cos \alpha, \sin \alpha]^\top$ is the direction of the rotation axis. The orthogonal vector $\mathbf{e} = [-\sin \alpha, \cos \alpha]^\top = \mathbf{a}^\perp$ is the epipolar direction. A straightforward computation shows that

$$\mathbf{M}\mathbf{e} = s \frac{f'}{f} (\cos \rho + \sin \rho (a \sin \alpha - b \cos \alpha)) \mathbf{e}, \quad (7)$$

thus giving an analytic proof of the following result:

Theorem 1. If the rigid motion between two weak-perspective cameras is assumed to be free of cyclorotation, then the epipolar direction \mathbf{e} can be recovered as one of the two eigenvectors of the vectorial part \mathbf{M} of the affinity that relates two views of a planar scene.

As a consequence, the direction $\mathbf{a} = \mathbf{e}^\perp$ of the axis of rotation can also be recovered.

Figure 1 illustrates the above result. Two views \mathcal{R} and \mathcal{R}' of a planar H-shaped object are shown, which are related by a rotation about an axis parallel to the image plane (i.e., free of cyclorotation). For simplicity of illustration, a basis $\{\mathbf{r}_1, \mathbf{r}_2\}$ is chosen aligned with the main axes of the H, and the axis of rotation is taken to be parallel to \mathbf{r}_2 . Thus, the gray plane swept by \mathbf{r}_1 is left invariant by the rotation. Note, then, that the epipolar direction is that of \mathbf{r}_1 in \mathcal{R} and that of $\mathbf{M}\mathbf{r}_1$ in \mathcal{R}' , and its perpendicular within each image is the direction of the rotation axis.

A geometric proof of Theorem 1 is included in [11]. Within the same geometrical framework, this result is generalized to the affine camera model leading to Theorem 2. Let us sketch the main ideas of this generalized result; the reader is referred to [11] for the details of the proof. The main advantage of this generalization is that, within the affine camera model, the projected target does not need to be centered in the image (assuming that the image center is a good approximation to the principal point). This enables us to handle a broader range of situations where the condition of small field of view is satisfied but the condition of being centered is relaxed. The affine camera model, which encloses the weak-perspective one, projects a scene point first under a fixed direction (which corresponds to a point \bar{O} lying on the plane at infinity Π_∞) onto the average

depth plane \mathcal{R}^C (the plane parallel to the image plane \mathcal{R} containing the centroid C of the scene object), and then perspectively from this fronto-parallel plane \mathcal{R}^C onto the image \mathcal{R} . When \overline{O} equals the direction O orthogonal to the image plane, the affine camera becomes a weak-perspective camera. By this projection procedure it is inferred that the affine camera, as well as the weak-perspective camera, preserves parallelism.

While in the weak-perspective camera model the improper optical center O is determined by the orientation of the image plane (i.e., O is the pole with respect to the absolute conic Ω of the improper line r of \mathcal{R}), in the affine camera model the improper optical center \overline{O} may be any point in Π_∞ . In fact, the direction of parallel projection, i.e., the improper optical center, depends on the position of the projected target within the image plane. This implies, on the one hand, that the same (pinhole) camera under affine viewing conditions can take two affine views with different improper optical centers (but keeping the same image plane). On the other hand, this also implies that, while the orientation of the image plane (and hence the improper optical center in case of a weak-perspective camera) is determined by the displacement performed by the camera, the improper optical center is not determined by the camera motion in the more general case of an affine camera. This is one of the reasons that makes the affine camera model more difficult to handle than the weak-perspective one.

Since the improper optical centers lie at infinity, the epipoles (of the first and second affine cameras) are also located at infinity in the image planes, i.e., the epipolar lines in both views are parallel. But, while in the weak-perspective cameras the epipoles coincide with the orthogonal direction (in the image plane) of the axis of rotation, in the general affine cameras the epipoles are no more related to this distinguished direction and, thus, a priori, they do not provide information about the rigid motion between the two affine cameras. This explains why most of the literature about the general affine camera model switches to the weak-perspective camera model when the question of inferring camera motion is addressed. Let us state the announced generalization result:

Theorem 2. Assume that the rigid motion between two affine cameras is free of cyclorotation and that the target projections are shifted (from the center of the image) along the direction orthogonal to the axis of rotation. Then the epipolar direction can be recovered as one of the two eigenvectors of the vectorial part \mathbf{M} of the affinity that relates the two affine views of a planar scene.

2.3 Computing the Epipolar Direction from the Affinity Parameters

Fix any coordinate frame in the image (for instance pixel coordinates, since orthonormality is not required) and assume that the affinity that relates the two views has the expression

$$\mathbf{x}' = \mathbf{M}\mathbf{x} + \mathbf{t} = \begin{bmatrix} M_{1,1} & M_{1,2} \\ M_{2,1} & M_{2,2} \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix} + \begin{bmatrix} t_x \\ t_y \end{bmatrix}. \quad (8)$$

In virtue of Theorem 1, the epipolar direction is one of the eigenvectors of \mathbf{M} . An eigenvector $[1, w]^\top$ of \mathbf{M} satisfies the equation

$$M_{1,2}w^2 + (M_{1,1} - M_{2,2})w - M_{2,1} = 0. \quad (9)$$

If the motion is under the hypothesis of Theorem 1, then (9) must have two real solutions w_1, w_2 , and the epipolar direction is $\mathbf{e} = [1, w_i]^\top$, for some $i \in \{1, 2\}$ (or $[0, 1]^\top$, in case $M_{1,2} = 0$).

3 Extracting the Affinity Parameters in Our Implementation

The affinity that relates two affine views is usually computed from a set of point matches. However, point matching is still one of the key bottlenecks in computer vision. In this work an active contour [12] is used instead. The active contour is fitted to a target object and the change of the active contour between different views is described by a shape vector deduced as follows. The contour is first represented as a parametric spline curve as it is common in Computer Graphics [13]. It has previously been shown [12] that the difference in control points $\mathbf{Q}' - \mathbf{Q}$ may be written as a linear combination of six vectors. Therefore, using matrix notation,

$$\mathbf{Q}' - \mathbf{Q} = \mathbf{W}\mathbf{S}, \quad (10)$$

where

$$\mathbf{W} = \left(\begin{bmatrix} 1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 1 \end{bmatrix}, \begin{bmatrix} \mathbf{Q}^x \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ \mathbf{Q}^y \end{bmatrix}, \begin{bmatrix} 0 \\ \mathbf{Q}^x \end{bmatrix}, \begin{bmatrix} \mathbf{Q}^y \\ 0 \end{bmatrix} \right), \quad (11)$$

and \mathbf{S} is a vector with the six parameters of the linear combination, the shape vector

$$\mathbf{S} = [t_x, t_y, M_{1,1} - 1, M_{2,2} - 1, M_{2,1}, M_{1,2}]^\top, \quad (12)$$

which encodes the relation between different affine views of the planar contour.

Note that the dimension of the shape vector can be reduced if robot motion is constrained, for instance to lie on a plane [14].

Once the compact representation of the contour in terms of control points and knots is obtained, a Kalman filter is used to track the contour along the sequence [12], and the shape vector is updated at each frame.

In previous works [15,16], the continuously updated shape vector was used to estimate robot egomotion in practice, provided data from other sensors (such as an inclinometer) or scene information (such as depth) were supplied. Here we focus on the extraction of epipolar direction from the shape vectors of just two views, and the analysis of the attainable accuracy in the different possible working conditions.

4 Experimentation

Two sets of experiments were performed to evaluate the accuracy of the proposed method. The first set uses synthetic image sequences generated by simul-

ating camera motion and computing projections under a full perspective camera model. Using this set, the sensitivity of the proposed algorithm to perspectivity effects is assessed by changing the distance of the target to the camera. A complete study involving the relaxation of all weak-perspective hypotheses can be found in [11].

The affine epipolar geometry is usually estimated using the Gold Standard algorithm [5]. This technique requires image correspondences of at least 4 non-coplanar points. Using also our synthetic experimental testbed, we show the effects of approaching coplanarity for this configuration, and compare the results with those of our method.

The second set of experiments uses real images taken by a robot arm moving along a calibrated path, showing the performance of the approach under realistic imaging conditions. In this setting, a comparison with the Gold Standard algorithm is also provided.

4.1 Simulations

When synthetic images are generated using an affine camera model (i.e., assuming perfect weak-perspective conditions), the epipolar direction is exactly recovered with the proposed method. However, we would like to assess the validity of the method under more general conditions. To this end, we generate the test set of synthetic images using a full perspective camera model. Then, of course, perspectivity effects affect the recovery of the epipolar direction in the ways that will be analysed in the following.

In the first experiment we analyse how a decrement of the distance Z_{ave} from the camera to the target affects the computation of the epipolar direction. Decreasing the distance enlarges perspective effects, and consequently, should increase the error in epipolar direction recovery. For this experiment we consider distances of 500, 750, 1000, 1250, 1500, 1750 and 2000mm. The smallest of these, 500mm, corresponds to an extreme situation for the weak-perspective model, in which important unmodelled distortions in the projected control polygon are present. For larger depth values, the affine conditions are better satisfied, thus reducing the error, as shown in Figure 2. It is worth noting that even under these unfavourable conditions the recovery error stays below 0.6° .

The effects of relaxing other assumptions, such as lateral translations leading to uncentered targets, introducing depth relief, or having cyclorotation have also been explored and the results are given in [11], where the sensitivity to contour shape is also analysed.

Next we describe a comparison with a standard technique for computing the affine epipolar geometry, namely the Gold Standard (GS) algorithm [5]. This algorithm, contrary to our procedure, needs non-coplanar point correspondences in order to compute the maximum likelihood estimate of the affine fundamental matrix. While in theory, only four non-coplanar points would suffice for computing the affine epipolar geometry using the GS algorithm, its performance is affected by the amount of non-coplanar information provided, both in terms of depth range and in the number of points used. The idea is to establish experimen-

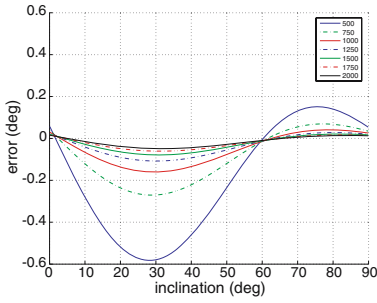


Fig. 2. Effects of relaxing one of the weak-perspective conditions by varying the distance from the camera to the target. The camera rotation is of 40° about an axis on the target with inclination of 45° .

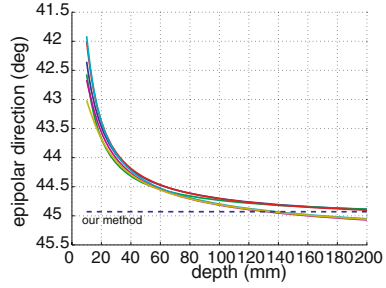


Fig. 3. Epipolar direction computed with the GS algorithm in the case of 2,4,...,12 out-of-plane points (a curve for each number) placed at increasing depths (in abscissae) above the H-shaped contour

tally the amount of depth information required by GS algorithm for it to provide equivalent epipolar direction recovery results to our procedure.

To this end, we set first an experiment in which we add a range from two to twelve extra points to the H-shaped contour, varying their distance with respect to the contour plane. Camera parameters are fixed at: 500 mm distance to target and a focal distance of 767 pixels. As before, camera motion is achieved via a rotation of 40° about an axis placed at an orientation of 45° on the target plane. The results are shown in Figure 3. It can be seen how as the depth of these points is increased, the error in the computation of the epipolar direction decreases. Moreover, it turns out that the number and xy location of these points have little effect in the computation of the epipolar direction. The figure contains plots of the resulting errors in the computation of the affine epipolar direction with the GS algorithm for different numbers of out-of-plane points, and a threshold indicating the error in the recovery of the epipolar direction using

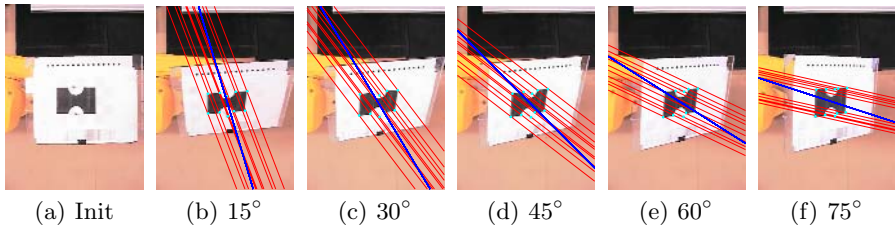


Fig. 4. The first experiment with real images entails pairs of views consisting of the initial one plus each of the other five, corresponding to camera rotations of 40° about an axis on the target with inclinations sampled at intervals of 15° . The epipolar direction computed by the proposed technique is displayed as a line passing through the target center, while the thin lines are the epipolar lines obtained with GS.

Table 1. Mean and standard deviation in degrees of the epipolar direction computed by the proposed technique and the GS algorithm from real images

epipolar direction	-15	-30	-45	-60	-75
$\bar{\theta}$	-16.63	-31.01	-45.00	-57.63	-72.04
σ	0.14	0.09	0.14	0.19	0.13
θ_{GS}	-18.53	-34.25	-49.46	-62.53	-76.36

our proposed technique under the same experimental conditions (the additional points out of the contour plane are evidently not used in this case). As shown in the figure, for the given experimental conditions, the results of our technique are comparable to those of the Gold Standard algorithm when the extra points are placed roughly at a distance equal to the target size (120 mm in our case).

Note the importance of parallax in the computation of the affine fundamental matrix with the Gold Standard algorithm. As the target points approach coplanarity, the parallax vector, which determines the epipolar direction, is monotonically reduced in length. Consequently, the accuracy of the line direction is also reduced, and the covariance of the estimated affine fundamental matrix increases. This situation does not occur in our procedure, as it has been devised precisely to compute the affine epipolar direction from two views of a plane.

4.2 Experiments Using Real Images

We present now results on image sequences in a controlled setting of our technique for computing the affine epipolar direction from pairs of views of a plane only. The goal of this work is not tracking, but computing the affinity from an active contour deformation, and using it to estimate the epipolar direction induced by the two views. To this end, we facilitate the tracking phase by moving a simple target placed on a manipulator end-effector, and focus on evaluating the accuracy of the direction recovered in different situations, compared to robot motion ground truth.

The experimentation setup consists of a Stäubli RX60 manipulator holding the target pattern on its end-effector. This target is a planar artificial H-shaped figure with corners and curved edges, which can be easily tracked with our active contour tracker. We are interested in using such setup in order to obtain a precise ground truth for the experiment. The initial distance from camera to target has had to be set to 500 mm. This corresponds to the extreme case discussed in Section 4.1, Fig. 2, and, therefore, we are testing the proposed approach under relaxed weak-perspective conditions. The acquired images have evident perspective effects, as shown in Figures 4 and 5, which make our algorithm work under extreme conditions. In order to provide depth information to the GS algorithm, the endpoints of two 20 mm screws placed at both sides of the contour are used as matching features in junction with the eight corners of the contour. Note that these are also extreme conditions for the GS algorithm to work, since

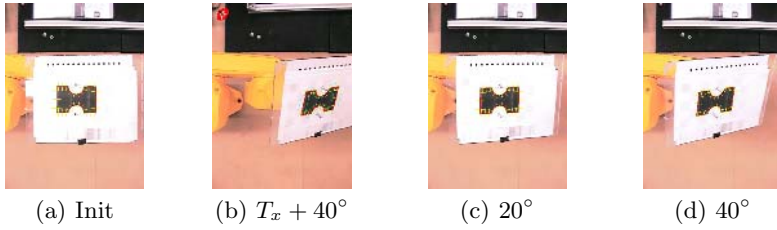


Fig. 5. Experiments with real images further relaxing weak-perspective conditions. The first sequence, entailing an uncentered target, starts at (a) and ends at (b). The next one departing from a non-frontoparallel target position starts at (c) and ends at (d).

Table 2. Mean and standard deviation of the epipolar direction computed over real images when weak-perspective conditions are further relaxed

<i>Frames</i>	$\bar{\theta}$	σ	θ_{GS}
Not Centered	-34.65	0.13	-56.29
Not Frontoparallel	-43.89	0.09	-49.78

very little depth information is provided: only two out-of-plane points. Thus, due to the setup we currently have, we are comparing both algorithms at the limit of their respective working conditions.

The first experiment entails camera motion induced by a rotation of 40° about an axis on the target at various inclination angles sampled at intervals of 15° . This, thus, relates to Fig. 2 with distance equal to 500 mm . Starting from the fronto-parallel position shown in Figure 4(a), the contour is tracked to each of the final views shown in the remaining frames of the figure. The epipolar direction computed by the proposed algorithm in each case is displayed as a line passing through the target center. Thin lines passing through the points correspond to the epipolar direction computed with the GS algorithm.

Table 1 presents the numerical values obtained in the computation of the epipolar direction. Standard deviation is computed by acquiring 300 images in the final position, estimating the shape vectors and then computing the corresponding epipolar directions. Note that the standard deviations are all very similar, and the mean values deviate more from ground truth as the angle departs from the 45° inclination. This should be interpreted in the light of Fig. 2 as meaning that the tracker amplifies the recovery error due to perspectivity effects unmodelled by the weak-perspective camera. Consequently, under true weak-perspective conditions, the errors should be much lower as indicated by the shrinking of the error curves in Fig. 2 when the distance Z_{ave} from the camera to the target increases. Results using the GS algorithm are slightly worse than those obtained with the proposed algorithm. This is due to perspective effects as well as to the poor depth information provided with the point matches used.

Two additional sequences were analyzed after further relaxing weak-perspective conditions. The first such sequence, labelled “Not centered”, starts at the

fronto-parallel initial position (Fig. 5(a)) and finishes at an uncentered position, after a translation of 100 *mm* along the *x* axis of the robot coordinate frame and a rotation of 40° about an axis at 45° inclination (Fig. 5(b)). Consistent with our simulated results [11], this lateral camera translation is by far the violation of weak-perspective conditions that has the most pervasive effect on the computation of the epipolar direction. See the numbers in Table 2, first row, which is far from the motion assumption of Theorem 2. This pervasive effect appears also in the computation with the GS algorithm, yielding the largest error in the experiments.

The second experiment, labelled “Not Frontoparallel”, corresponds to the same rotation described above, but the initial frame is not frontoparallel. The sequence starts with the target already rotated 20° as shown in Fig. 5(c) and, after a further rotation of 20° , finishes at 40° (Fig. 5(d)), all rotations about an axis at 45° inclination as before. Observe that the result is only a bit worse than that of the initial experiment, but with a similar standard deviation. The result with the GS algorithm here is similar as before.

5 Conclusions

The recovery of camera motion and scene structure from uncalibrated image sequences has received a lot of attention lately due to its numerous applications, which range from robot localization and navigation, to virtual reality and archeology, to name just a few. Most works rely on detecting a set of non-coplanar points in the scene and matching their projections on the different views. In this paper we have departed from this main stream, by dealing with a less informative situation, namely features lying on a plane, and recurring to contour tracking instead of point matching.

Our main result is that, under weak-perspective conditions and assuming a camera motion free of cyclorotation, the epipolar direction can be recovered from the affinity relating two views of a planar scene.

Synthetic images were used to evaluate the results in a noise-controlled environment, and then to compare the accuracy of our method with that of the Gold Standard algorithm, which relying on matches of non-coplanar points falls in the main stream mentioned above.

The outcome of the comparison has been very encouraging, since with less scene information (only from a plane) and with a much simpler processing (solving a single second-order equation), we are able to obtain the epipolar direction with similar accuracy. It is worth reminding, however, that our method is less general in that it requires a camera motion free of cyclorotation.

The second experimental set consisted of image sequences that were used to validate the proposed approach under real imaging conditions. Note that the objective of the paper is to show what can be obtained from the affine deformation of two views of a contour, and not to validate the robustness of the contour tracker used. For this reason, simple and well-calibrated image sequences were used in order to have a good basis for ground truth comparison.

Future work will include an error analysis that involves positional errors on the contours due to the image acquisition process. Moreover, we will try to unravel under what circumstances additional information on camera motion and scene structure can be recovered from two (or more) uncalibrated views of a planar object. Along the same line, we will tackle the recovery of the orientation of the scene plane, as well as what occurs in degenerate situations in which such orientation is the same as that of the image plane, or when both planes have a common direction.

References

1. Beardsley, P.A., Zisserman, A., Murray, D.W.: Sequential updating of projective and affine structure from motion. *Intl. J. of Computer Vision* **23** (1997) 235–259
2. McLauchlan, P.F., Murray, D.W.: A unifying framework for structure and motion recovery from image sequences. In: *Proc. Intl. Conf. on Computer Vision*. (1995) 314–320
3. Koenderink, J., van Doorn, A.J.: Affine structure from motion. *J. Opt. Soc. Am. A* **8** (1991) 377–385
4. Shapiro, L., Zisserman, A., Brady, M.: 3d motion recovery via affine epipolar geometry. *Intl. J. of Computer Vision* **16** (1995) 147–182
5. Hartley, R., Zisserman, A.: *Multiple View Geometry in Computer Vision* (Second Edition). Cambridge University Press (2004)
6. Sturm, P., Maybank, S.J.: On plane-based camera calibration: a general algorithm, singularities, applications. In: *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*. Volume 1. (1999) 432–437
7. Demirdjian, D., Zisserman, A., Horaud, R.: Stereo autocalibration from one plane. In: *Proc. 6th European Conf. on Computer Vision*. (2000) 625–639
8. Malis, E., Cipolla, R.: Camera self-calibration from unknown planar structures enforcing the multiview constraints between collineations. *IEEE Trans. on Pattern Analysis and Machine Intelligence* **24** (2002) 1268–1272
9. Bartoli, A., Sturm, P., Horaud, R.: Structure and motion from two uncalibrated views using points on planes. In: *Proc. 3rd. Intl. Conf. on 3D Digital Imaging and Modeling, Canada* (2001) 83–90
10. Kaminski, J.Y., Shashua, A.: On calibration and reconstruction from planar curves. In: *Proc. European Conf. on Computer Vision*. (2000) 678–694
11. Alberich-Carramiñana, M., Alenyà, G., Andrade-Cetto, J., Martínez, E., Torras, C.: Affine epipolar direction from two views of a planar contour. Technical Report IRI-DT-2005/03, Institute of Robotics (IRI) (2005)
12. Blake, A., Isard, M.: *Active Contours*. Springer (1998)
13. Foley, J., van Dam, A., Feiner, S., Hughes, F.: *Computer Graphics. Principles and Practice*. Addison-Wesley Publishing Company (1996)
14. Alenyà, G., Martínez, E., Torras, C.: Fusing visual and inertial sensing to recover robot egomotion. *Journal of Robotics Systems* **21** (2004) 23–32
15. Martínez, E., Torras, C.: Qualitative vision for the guidance of legged robots in unstructured environments. *Pattern Recognition* **34** (2001) 1585–1599
16. Martínez, E., Torras, C.: Contour-based 3d motion recovery while zooming. *Robotics and Autonomous Systems* **44** (2003) 219–227

Toward Visually Inferring the Underlying Causal Mechanism in a Traffic-Light-Controlled Crossroads

Joaquín Salas, Sandra Canchola, Pedro Martínez,
Hugo Jiménez, and Reynaldo C. Pless

CICATA Querétaro, Instituto Politécnico Nacional,
Querétaro, Mexico, 76040

Abstract. The analysis of the events taking place in a crossroads offers the opportunity to avoid harmful situations and the potential to increase traffic efficiency in modern urban areas. This paper presents an automatic visual system that reasons about the moving vehicles being observed and extracts high-level information, useful for traffic monitoring and detection of unusual activity. Initially, moving objects are detected using an adaptive background image model. Then, the vehicles are tracked down by an iterative method where the features being tracked are updated frame by frame. Next, paths are packed into routes using a similarity measure and a sequential clustering algorithm. Finally, the crossroads activity is organized into states representing the underlying mechanism that causes the type of motion being detected. We present the experimental evidence that suggests that the framework may prove to be useful as a tool to monitor traffic-light-controlled crossroads.

1 Introduction

Nowadays, the protection of people and property is a major concern. This has resulted in a large number of cameras being installed to perform surveillance and monitoring tasks in places such as banks, offices, houses, and streets. The question which now arises is what will it be the best method to process all this information. While there have been numerous cases where the video footage was helpful in the retrospective analysis of an event, nonetheless, as [3] point out, all this stream of information may have better uses than just as a forensic tool to reconstruct past events. Although stored video may prove helpful in many situations, it is worth trying to use video monitoring systems as proactive tools that provide information which may prove useful in preventing harm.

In our research, we aim to develop automatic image analysis tools with the capacity to infer high-level descriptions of the scene being monitored. This is an area where one can expect major advances in the near future. It may require the interpretation of long spatio-temporal image sequences by agents focused on particular actions or multiple agents interacting. It seems that Hidden Markov Models (HMM) are particularly suited for this task. For instance, Oliver *et al.*

used them to detect and classify interactions between people[16]. Interesting interactions include following people and unusual changes in trajectories. Also, Wada and Matsuyama[24] use HMM to recognize the behavior of multiple moving objects. In their approach, feasible assumptions about the present behavior consistent with the input image and behavior models are dynamically generated and verified. In our work, we recognize the importance of computing the relation, and the strength, between the states that cause the motions and the trajectories being observed. The task of delivering high-level descriptions seems to demand the construction of elaborate abstractions. In this vein, Mark and Ellis[13] group together paths into routes and then give semantic meaning to motion activities within the routes. A possible way to represent this hierarchy of knowledge is through the use of graph or tree data structures. For instance, Lou *et al.*[11] represent the set of routes as the root. Then, the branches may represent subsets of routes. At the leaves, they represent the individual paths. This is very similar to the model that we use. However, since our focus is on the structured scenario of periodic changes existing in a traffic-light-controlled crossroads, our model gravitates around the different states of activity.

In the present study, we center our attention on the identification of the set of routes that are present in each state of activity in a crossroads. Studying what happens in this type of location has important implications for modern urban areas. For instance, as Mussonne and Sala[15] point out, up to one third of the collisions take place at crossroads. In our model, the type of motions that we observe represent meaningful states that have a causal function, *i.e.*, a particular configuration of traffic light motivates certain trajectories while inhibiting others. This cause-effect relationship repeats itself into cycles of activity. There have been numerous studies for specific types of situations that arise at vehicular intersections, like incident detection [6], vehicle classification [7], vehicle counting [14], and vehicle speed [4]. In this paper, we advance a methodology to identify the different states that constitute a process.

The rest of the document develops as follow. In §2, we discuss how a moving object is tracked to compute a vehicle path. Then, in §3, paths are packed together into routes using a similarity measure. Next, in §4, the activity in the crossroads is organized into the states provided by the traffic-light combinations. In §5, we provide some results from experimentation. Finally, we present our conclusion.

2 Computing a Vehicle's Path

Reliably computing vehicle trajectories is an important step which forms the basis for the rest of the analysis. It implies detecting the moving elements and tracking them along their trajectory.

2.1 Detecting Moving Vehicles

In this study, we center our attention on fixed cameras looking at dynamic scenes where there are some objects that remain static and some others that move. A

primary tool to extract information about moving objects is background subtraction [19]. However, it has become clear that for vision systems to work for extended periods of time, the background model should be updated dynamically. In a seminal work, Stauffer and Grimson[20] suggested to use a mixture of Gaussian distributions to describe the changes in the dynamic behavior of a pixel. Nonetheless, in particular for an object passing in front of another, it is not clear that the variations observed can be modeled well by a parametric distribution. This was recognized by Elgammal *et al.*[5]. Other researchers, like Rigoll *et al.*[17], used adaptive state estimation in the form of Kalman filters to obtain reliable results even in cases with considerable process variation and observation uncertainties.

In the context of stereovision, Tomasi and Manduchi[22] proposed to characterize the lines of each individual image in terms of both its intensity and gradient. They noted that in the presence of occlusions corresponding scanlines showed an obvious deviation with respect to one another. In [18], this observation is made effective by detecting when, in the temporal axis, both accretions and occlusions occur. In general, a particular pixel value that belongs to a static background tends to stabilize around a certain value. The differences with respect to this value can be described in terms of a statistical model. However, background occlusion due to a foreground object passing by is a random process that is difficult to describe using a parametric model. Under these circumstances, the phenomenon observed seems to be present in two forms. In one of them, the values remain stable until an object occludes the background. When the object has passed by, the trajectory described by the curve returns to its attraction point. A second case is when the foreground objects integrates into the background. In that case, the attraction point assumes a new position from then on.

2.2 Tracking Vehicles

Correspondence is a basic problem that has received much attention from researchers. For instance, Coifman[2] proposed a method to track vehicles where



Fig. 1. The camera is placed on top of this 28-meter tower

occlusion may be present. In a seminal work, Lucas and Kanade[12] studied the problem of tracking a bidimensional feature over consecutive frames using a Newton-Raphson type of technique. In their formulation, the feature experiences transformations such as translations. In principle, it is possible to include a search for rotations. However, as Tomasi and Shi[23] observed, this may be error prone. In the present study, once the feature under consideration has been found it is replaced by the current feature view. On the one hand, if the frames are close enough to each other in time, the feature will not change excessively. On the other, updating the feature after each frame allows to the system to handle changes in the vehicles' appearance beyond bi-dimensional image transformations.

3 From Paths to Routes

Paths that are similar to each other can be packed into routes. This is done throughout a non-supervised clustering algorithm [21]. That is, the labeling for the set of paths used for learning is not available. Hence, the prime problem is to unveil the organization pattern of the routes with the aim to obtain relevant conclusions. In our case, let $X = \{x_1, \dots, x_a\}$, the set of trajectories. Our objective is to define m clusters C_1, \dots, C_m such that the following conditions are met

- $C_i \neq \emptyset, i = 1, \dots, m$; there is no empty set.
- $\bigcup_{i=1}^m C_i = X$; the set of routes is the set of paths.
- $C_i \cap C_j = \emptyset, i \neq j; i, j = 1, \dots, m$; the routes do not have paths in common.

In order to compute the similarity between paths, we use the Hausdorff distance. Once similarity is computed, a sequential clustering algorithm is used.

3.1 Dissimilarity Between Paths

The Hausdorff distance allows us to compare two sets of points. It has been used with success in applications as diverse as face recognition [1] and people localization within an image [8]. Among other advantages, the Hausdorff distance does not require the cardinality of the sets to be the same. Indeed, it has been shown that the algorithm complexity of the direct implementation can be reduced by interpreting it in the Voronoi space of the points layout [9]. In this manner, given two sets of points $A = \{\mathbf{a}_1, \dots, \mathbf{a}_p\}$ and $B = \{\mathbf{b}_1, \dots, \mathbf{b}_q\}$, the Hausdorff distance is defined as

$$H(A, B) = \max(h(A, B), h(B, A)), \quad (1)$$

with

$$h(A, B) = \max_{\mathbf{a}_i \in A} \min_{\mathbf{b}_j \in B} \rho(\mathbf{a}_i, \mathbf{b}_j). \quad (2)$$

The above equation is known as the direct Hausdorff distance. There, $\rho(:, :)$ measures the distance between two points. This is similar to the approach followed

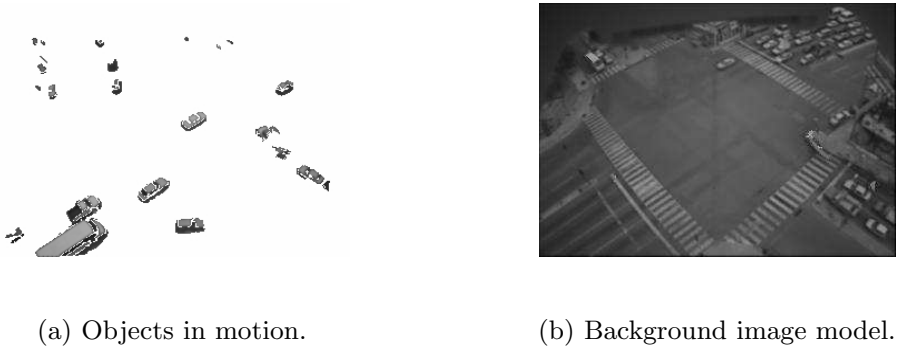


Fig. 2. Building the background model

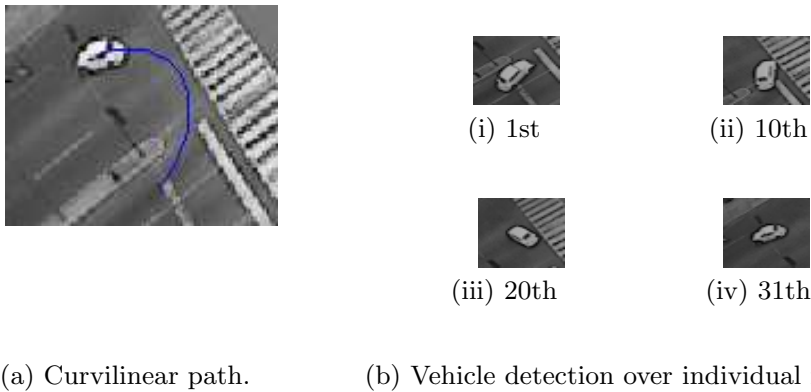


Fig. 3. Tracking vehicle paths. The sought feature is refreshed after each iteration, making it possible to adjust for changes in appearance.

by Junejo *et al.*[10]. They represent a trajectory by a node in a graph. Each node is connected with each other and the weight of an edge is the distance of comparing two trajectories using the Hausdorff distance. In this study, we are using the Euclidian distance defined as

$$\rho(\mathbf{a}, \mathbf{b}) = \| \mathbf{a} - \mathbf{b} \| . \tag{3}$$

3.2 Inferring Routes

It is assumed that the number of routes is not known in advance. The basic idea is to find those paths that are similar to each other. Thus, given $X = \{x_1, \dots, x_n\}$, the set of paths, the objective is to compute the set of routes $C = \{C_1, \dots, C_m\}$. We start this by computing $D_{n \times n} = \{H(x_i, x_j)\}$, a matrix that express the dissimilarity between paths x_i and x_j , as dictated by Eq. (1).

Firstly, every path is initially set as non-visited. So for every non-visited path x_p , we mark it as visited and compute the set $U = \{u_1, \dots, u_s\}$ of non-visited

paths such that the dissimilarity measure, $H(x_i, u_k)$ for $k = 1 \dots, s$, is less than a predefined threshold γ . As a result, the route C_t is formed by the union of x_p and the paths in U . This procedure of selection and expansion is repeated for every member of U and subsequent adjoint sets are combined into C_t . The route is defined when no more paths can be incorporated. If there are non-visited paths then there is space for another route and the procedure should be repeated again.

4 Crossroads Activity

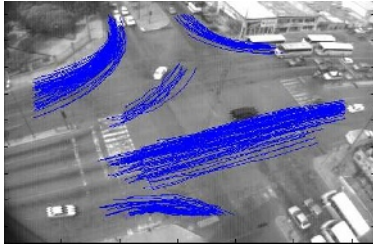
Now, the problem is to organize routes into states of a process defined by the traffic-light changes. That is, our task is to organize the routes based on the limits imposed by the traffic lights. Each path has associated a time stamp for each of the frames where the object was detected. We term the routes *exclusive* and *inclusive* depending on whether they are present only in a certain state or in several of them, respectively. In principle, *exclusive* routes may be uncovered by a frequency analysis because no matter how long they last, they repeat periodically. Subsequently, we may use these routes as a base to organize the rest of them. However, in the process we may miss important information about the transition between observations and states. Let $S_{n \times n}$ be an adjacency matrix whose rows and columns are the discovered routes. As time passes, $S(i, j)$ increases by one each time a transition between route C_i and route C_j occurs. That is, if a path belonging to C_i was detected at time k and a path belonging to C_j was detected at time $k+1$, then $S(i, j)$ is increased by one. The adjacency matrix S becomes the transition matrix (in HMM terminology), T , by computing the sum over the elements in a particular row and then dividing each element by that sum. Thus, T represents, in its rows, the states of the process and, in its columns, the routes that appear in it. The next step is to identify which one corresponds to the underlying causal mechanism. Suppose that we know N , the number of independent states in the process. In most cases, this is a trivial number to obtain because it corresponds to the different traffic-light combinations that occur in an intersection. What we want is to compute the combination of N rows of T , from the n available, that correspond to the independent states of the process. Let $\mathcal{T} = \{1, 2, \dots, n\}$ be the enumeration of states in T and let $R = \{r_1, \dots, r_N\}$ be a set of N numbers such that $r_i \in \mathcal{T}$, we want to estimate the set of N numbers of \mathcal{T} that are the indices to the states that maximize the following expression

$$q = \sum_{i=1}^n \max_{r_j \in R} T(r_j, i). \quad (4)$$

At the end, the set R associated with the maximum q corresponds to the indices of the states of the routes of the transition matrix.

5 Experimental Results

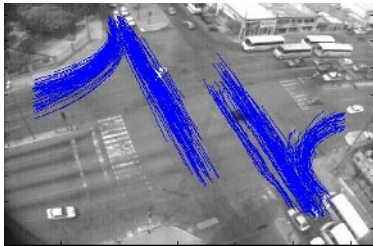
The exposed model was implemented through the development of diverse visual routines. For experimentation a set of 20,000 images with a resolution of 320



(a) Motion from left to right.



(b) Motion from right to left.



(c) Motion up and down and vice versa.

Fig. 4. Paths organized by states

rows times 240 columns was used. In our data set, there are approximately 10 complete traffic-light cycles . Each cycle is composed of three states.

The moving-object detection algorithm processes about five frames a second. When an object stops for about ten frames, it is incorporated as part of the image background model. The procedure has been tested under diverse weather and illumination conditions, including cloudy, sunny, and rainy days. For all conditions, the operational parameters were maintained unchanged. Fig. 2 shows a typical example of the current image, the background model, and the moving objects detected.

Path detection is largely based on the motion-detection stage. When an object in motion is detected, a small 11 x 11 pixel window around the object centroid is selected. This window is tracked until it is missed. This normally occurs when the feature leaves the camera’s field of view. In each frame, the window’s contents is refreshed with the current object appearance. After tracking down the object, there is a set $P = \{\mathbf{p}_1, \dots, \mathbf{p}_n\}$ containing information about the position and time of the detected object. In Fig. 3, we show an example of the path traced by a vehicle performing a U-turn. This example is important since it shows that features are tracked successfully even when the object appearance changes between the beginning and the end of its path. As long as the time span between images is not too long and the velocity of the object is not too large, the tracking algorithm will work satisfactorily.

Similar paths are clustered together into routes, using the algorithm described previously. Only paths larger than 40 pixels are considered. Two paths are packed together whenever their dissimilarity, using the Hausdorff distance, is below a certain threshold. After experimenting, we arrive at a value of $\gamma = 30$ pixels. For the 20,000 frames, 1,305 paths were detected. Thereafter, only routes with more than 10 paths were considered. We obtained 18 routes out of the 1,305 paths.

In fact, each path detected has time stamps associated with all the points along its trajectory. So at this point, we can organize how paths occur throughout time and we can organize this information by the route to which each path belongs. From these path-route-time relationships, we build the transition matrix T that represents the possible states of the process. Using the *a priori* knowledge that there are three states, we maximize Eq. (4) to select which states of T are more exclusive. This gives the result shown in Fig. 4. Perhaps, a most important piece of information may be the transition matrix T . It expresses the observed frequency of transition between routes.

6 Conclusion

In this document, we advance toward the construction of a visual system capable of automatically inferring the activity causal underlying control mechanism. In order to study the problem, gather data and unveil the prime variables, a structured scenario, such as the light-controlled vehicular intersection, provides a rich testbed for analysis and development. In this document, a strategy to assign the observed activities to existing states is developed. The model adapts to changes in lighting conditions. Consistent tracking is achieved by refreshing the object being observed in each new frame. Tracking works even in the presence of L-turns or U-turns. Paths are packed into routes which in turn are analyzed based on their time occurrence to determine the set of them which make up each state.

In this high-level view, the process is divided into states. Each state is integrated with non-exclusive routes, which in turn are made up from paths. In the process, the model learns routes and assigns occurrence probabilities to each of them. The results presented here pertain to a well structured process. In the future, we plan to study other environments where the underlying causal mechanism presents more subtle facets.

Acknowledgments

This study was partially funded by CGPI-IPN. The authors thank CONCYTEQ and the Centro Educativo “Manuel Gómez Morín” for their help and support

References

- [1] Guo Baofeng, Lam Kin-Man, Lin Kwan-Ho, and Siu Wan-Chin. Human Face Recognition based on Spatially Weighted Hausdorff Distance. *Pattern Recognition Letters*, 24(1-3):499–507, 2003.

- [2] B. Coifman, D. Beymer, P. McLauchlan, and J. Malik. A Real-Time Computer Vision System for Vehicle Tracking and Traffic Surveillance. *Transportation Research*, 6(4):271–288, 1998.
- [3] R.T. Collins, A.J. Lipton, and T. Kanade. Introduction to the Special Section on Video Surveillance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):745–746, 2000.
- [4] D.J. Dailey, F.W. Cathey, and S. Pumrin. An Algorithm to Estimate Mean Traffic Speed using Uncalibrated Cameras. *IEEE Transactions on Intelligent Transportation Systems*, 1(2):98 – 107, 2000.
- [5] A. Elgammal, R. Duraiswami, D. Harwood, and L.S. Davis. Background and Foreground Modeling using Nonparametric Kernel Density Estimation for Visual Surveillance. *Proceedings of the IEEE*, 90(7):1151 – 1163, 2002.
- [6] G.L. Foresti and B. Pani. Monitoring Motorway Infrastructures for Detection of Dangerous Events. In *International Conference on Image Analysis and Processing*, pages 1144 – 1147, 1999.
- [7] S. Gupte, O. Masoud, R.F.K. Martin, and N.P. Papanikolopoulos. Detection and Classification of Vehicles. *IEEE Transactions on Intelligent Transportation Systems*, 3(1):37 – 47, 2002.
- [8] D. P. Huttenlocher, G. A. Klanderman, and W. J. Rucklidge. Comparing Images using the Hausdorff Distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 15(9):850–863, 1993.
- [9] Daniel P. Huttenlocher, Klara Kedem, and Micha Sharir. The Upper Envelope of Voronoi Surfaces and its Applications. In *Annual Symposium on Computational Geometry*, pages 194–203, 1991.
- [10] Imran Junejo, Omar Javed, and Mubarak Shah. Multi feature path modeling for video surveillance. In *IEEE International Conference on Pattern Recognition*, volume 2, pages 716 – 719, 2004.
- [11] Jianguo Lou, Qifeng Liu, Tieniu Tan, and Weiming Hu. Semantic Interpretation of Object Activities in a Surveillance System. *IEEE International Conference on Pattern Recognition*, 3:777 – 780, 2002.
- [12] B.D. Lucas and T. Kanade. An Iterative Image Registration Technique with an Application to Stereo Vision. In *Proceedings of the 7th International Joint Conference on Artificial Intelligence*, pages 674–679, 1981.
- [13] Dimitrios Makris and Tim Ellis. Path Detection in Video Surveillance. *Image and Vision Computing Journal*, 20(12):895–903, 2002.
- [14] O. Masoud, N.P. Papanikolopoulos, and E. Kwon. The Use of Computer Vision in Monitoring Weaving Sections. *IEEE Transactions on Intelligent Transportation Systems*, 2(1):18 – 25, 2001.
- [15] L. Mussone and G. Sala. An Analysis of Lateral Support Systems to Increase Safety at Crossroads. In *IEEE Intelligent Vehicles Symposium*, pages 383 – 388, 2003.
- [16] N.M. Oliver, B. Rosario, and A.P. Pentland. A Bayesian Computer Vision System for Modeling Human Interactions. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(8):831 –843, 2000.
- [17] G. Rigoll, S. Eickeler, and S.; Muller. Person tracking in real-world scenarios using statistical methods. In *Fourth IEEE International Conference on Automatic Face and Gesture Recognition*, pages 207–213, 2000.
- [18] Joaquin Salas, Pedro Martinez, and Jordi Gonzalez. Background Updating with the use of Intrinsic Curves . In *to appear in ICIAR and Lecture Notes in Computer Science*. Springer-Verlag, 2006.

- [19] M. Seki, H. Fujiwara, and K. Sumi. A robust background subtraction method for changing background. In *WACV00*, pages 207–213, 2000.
- [20] Chris Stauffer and W. E. L. Grimson. Adaptive Background Mixture Models for Real-Time Tracking. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 246–252, 1999.
- [21] Sergios Theodoris and Konstantinos Koutroumbas. *Pattern Recognition*. Academic Press, 1999.
- [22] Carlo Tomasi and Roberto Manduchi. Stereo matching as a nearest-neighbor problem. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):333–340, 1998.
- [23] Carlo Tomasi and Jianbo Shi. Good Features to Track. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 593–600, 1994.
- [24] T. Wada and T. Matsuyama. Multiobject Behavior Recognition by Event Driven Selective Attention Method. *Pattern Analysis and Machine Intelligence*, 22(8):873–887, 2000.

Computer Vision Based Travel Aid for the Blind Crossing Roads

Tadayoshi Shioyama

Department of Mechanical and System Engineering
Kyoto Institute of Technology
Matsugasaki, Sakyo-ku, Kyoto 606-8585, Japan
shioyama@kit.ac.jp

Abstract. This paper proposes a method for detecting frontal pedestrian crossings and estimating its length from image data obtained with a single camera as a travel aid for the blind. It is important for the blind to know whether or not a frontal area is a crossing. The existence of a crossing is detected in two steps. In the first step, feature points for a crossing is extracted using Fisher criterion. In the second step, the existence of a crossing is detected by checking on the periodicity of white stripes on the road using projective invariants. Next, we propose a method for estimating crossing length using extracted feature points. From the experimental results for evaluation, it is found that the existence of a crossing is successfully detected for all 173 real images which include 100 images with crossings and 73 images without crossing. The rms error of crossing length estimation for the 100 images is found 2.29m.

1 Introduction

An effective navigation system is an ultimate necessity to improve the mobility of millions of blind people all over the world. This paper addresses an application of computer vision as a travel aid for the blind. Since the range of detection of obstacles using a white cane, which is a usual travel aid for the blind, is very narrow, various devices have been developed such as the SONICGUIDE[1], the Mowat sensor[2], the Laser cane[3] and the Navbelt[4]. However, these devices are not able to assist the blind at a pedestrian crossing where information about the existence of a crossing and its length are important. There is a type of traffic light with particular equipment which notifies the blind of a safe direction at a crossing by sounds. An equipment using infrared[5] has been developed. The equipment is installed around the intersection to provide visually impaired with voice information and directions to go. But such equipments are not available at every crossing and it would take too long for such equipment to be put in everywhere. The "vOICe"[6] is commercially available which is a vision based travel aid for the blind using conversion of an image pixel information to sound. Although it can recognize walls, doors etc., its technique does not include the detection of pedestrian crossing.

Stephen Se[7] first addressed a pedestrian crossing detection by grouping lines in an image and checking on concurrency using the vanishing point constraint. However, a thorough evaluation of this technique is not performed yet and also the technique is working slow and far from real time.

We have aimed to develop a device with which the blind would be able to autonomously detect important information for safely negotiating a pedestrian crossing. In our previous paper[8] for the purpose of achieving such an objective using image data observed with a single camera, we proposed a method for image analysis of a crossing to measure the crossing length and to detect the state of the traffic lights. That method is based on an assumption that the area in front of the blind is a crossing. There has been a residual problem to detect whether or not the area in front of the blind is a crossing. In our previous paper[9] we proposed a method for detecting the existence of a crossing in the frontal area of the blind using image data observed with a single camera. The process of detecting the existence of a crossing is a pre-process followed by the process for measuring its length. In this paper we propose a unified method for detecting a crossing and measuring crossing length based on a new idea.

In the present method, at first, feature points for a crossing are extracted by using Fisher criterion. Next, the existence of a crossing is detected by checking on the periodicity of white stripes painted on the road using projective invariants[10] constructed from the feature points. The projective invariants are invariant even when the geometric shape of crossing observed in an image varies due to change in viewing direction. After detecting a pedestrian crossing, we find the feature point at the half crossing length, estimate a camera rotation angle by using four feature points at the half crossing length and measure the crossing length by using number of white bands between the nearest feature point and the feature point at the half crossing length.

In our previous paper[11], we developed a method for measuring the length of crossing by using number of white bands on a road and the band width, to improve the method of our previous paper[8] which is complicated, computationally inefficient, and needs many parameters to adjust. However it is based on an assumption that an optical axis of a camera is horizontal. In the present paper, we develop the method for measuring the length in the case where the optical axis is not always horizontal.

In order to evaluate the performance of the present method, experimental results are presented for detection of crossing and measurement of its length from real image data.

2 Detection of Pedestrian Crossing

We detect a pedestrian crossing using projective invariant derived from four colinear feature points. In this section, we represent the principle and method of pedestrian crossing detection.

2.1 Projective Invariant

We consider a one-dimensional coordinate x on a straight line in a three-dimensional (3D) space. We denote by \tilde{x} a homogeneous coordinate $(x, 1)^t$ for a coordinate x where the symbol “ t ” denotes the transpose operator. When we observe a point with homogeneous coordinate \tilde{x} on an image plane, the point \tilde{x} can be written according to a projective transformation[10] as follows:

$$\tilde{x}' = \lambda(\tilde{x})T\tilde{x}, \quad (1)$$

where T is a 2×2 dimensional matrix and $\lambda(\tilde{\mathbf{x}})$ is a parameter depending on $\tilde{\mathbf{x}}$. When we consider four colinear point $\mathbf{x}_i, i = 1, 2, 3, 4$, as illustrated in Fig.1, the Euclidean distance ℓ_{ij} between \mathbf{x}_i and \mathbf{x}_j is given by

$$\ell_{ij} = | \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_j |, \tag{2}$$

here $| \tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_j |$ is a determinant of a matrix $[\tilde{\mathbf{x}}_i \tilde{\mathbf{x}}_j]$ which is constructed with two column vectors $\tilde{\mathbf{x}}_i$ and $\tilde{\mathbf{x}}_j$. Hence under a projective transformation, ℓ_{ij} is transformed in the following:

$$\ell'_{ij} = \lambda(\tilde{\mathbf{x}}_i)\lambda(\tilde{\mathbf{x}}_j) | T | \ell_{ij}. \tag{3}$$

The cross-ratio of the Euclidean distances ℓ_{ij} :

$$I \equiv \frac{\ell_{12}\ell_{34}}{\ell_{13}\ell_{24}} \tag{4}$$

is invariant under the projective transformation because all $\lambda(\tilde{\mathbf{x}}_i), i = 1, 2, 3, 4$, and $| T |$ are cancelled out[10].

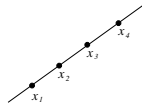


Fig. 1. Four collinear points

2.2 Pedestrian Crossing Detection with Projective Invariant[9]

In a pedestrian crossing, there are periodical white stripes, which are painted on a road surface. Inter-white striped distance is the same as the band width of a white stripe. Let b be the band width of a white stripe. We consider feature points, which are edge points of white stripes on a road surface. For four consecutive collinear feature points as illustrated in Fig.2, the projective invariant I of equation (4) is given by[10]

$$I = \frac{\ell_{12}\ell_{34}}{\ell_{13}\ell_{24}} = \frac{b \cdot b}{2b \cdot 2b} = 0.25. \tag{5}$$

The value of I is constant for four colinear consecutive feature points on a line with any direction. Hence the value of I obtained from four consecutive feature points in an observed image is equal to the value of I in equation (5). By using this invariant I , we can detect a pedestrian crossing in an observed image.

2.3 Feature Point Extraction by Fisher Criterion

At a feature point on an edge line of white stripe, which is usually observed as almost horizontal line, we set a vertical window with size of $(2 \times win + 1, 1)$ whose center coincides with a considered point as illustrated in Fig.3. The win is the size of the half window in pixel. The class 1 is defined as a set of intensities of grayscale image at pixels

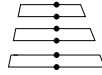


Fig. 2. Feature points in a crossing

in the upper side of the window. The class 2 is defined in the same manner as the class 1 for the lower side of the window.

The Fisher criterion fc is defined as the ratio of a “between class variance” var_b to a “within class variance” var_w . Maximizing fc implies maximizing the distance between the means of the two classes while minimizing the variance within each class:

$$fc = var_b / var_w, \tag{6}$$

where

$$var_b = p_1 p_2 (m_1 - m_2)^2,$$

$$var_w = p_1 var_1 + p_2 var_2,$$

m_i = the mean intensity in the class $i, i = 1, 2,$

var_i = the variance of intensity in the class $i, i = 1, 2,$

p_i = the probability of class $i, i = 1, 2.$

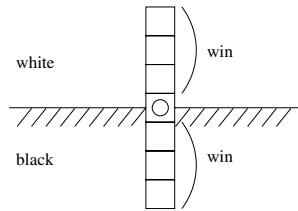


Fig. 3. Window at a feature point

The mean $m_i, i = 1, 2,$ has a high value if class i is corresponding to white region, and it has a low value if the class is corresponding to black region. When the center of the window coincides with a feature point, the Fisher criterion fc is considered to be high value. On the other hand, when the center of the window is different from a feature point, the fc is considered to be low value. The Fisher criterion fc is assigned to the center pixel of the window. By scanning the window in vertical direction, we can find feature points with local maximum fc .

2.4 Method for Extracting Feature Points

At first, we search two adjacent parallel vertical lines at the center of an image for feature points as illustrated in Fig.4. If we cannot find appropriate feature points, we search another left or right two adjacent parallel vertical lines whose x-coordinate are



Fig. 4. The extracted feature points



Fig. 5. The extracted feature points on left line

$w/4$ or $3w/4$, respectively, for feature points where w is the x-size of an image. Fig.5 shows the extracted feature points on the center and left lines, where the feature points on the left line are searched because the extracted feature points on the center lines do not satisfy the conditions (7), (8) for detection of crossing, that is, the projective invariants calculated from these extracted feature points are different from the projective invariant for crossing.

2.5 Method for Detecting Pedestrian Crossing

From n extracted feature points, we can calculate $(n - 3)$ projective invariants by equation (4). For each projective invariant $I(i), i = 1, 2, \dots, n - 3$, we check on whether it satisfies the following condition or not:

$$| I(i) - 0.25 | < 0.1 \times 0.25. \tag{7}$$

Denote by cnt the number of invariants satisfying the condition (7). We decide that there is a crossing if the following condition is satisfied:

$$cnt \geq \max\{[(n_c - 3)/5], 1\}, \tag{8}$$

where $[x]$ denotes the maximum integer less than x and $\max\{a_1, a_2\}$ denotes the larger value among a_1 and a_2 . n_c denotes the number of feature points from the nearest feature point to the one at the half crossing length in the image.

On detection for crossing, we extract feature points on two adjacent parallel vertical lines. We denote by $I_1(i), i = 1, \dots, n_1, I_2(i), i = 1, \dots, n_2$ the projective invariants calculated from the extracted feature points on the two adjacent parallel vertical lines. Let cnt_k and $n_{kc}, k = 1, 2$ be cnt and n_c for $I_k(i), i = 1, \dots, n_k, k = 1, 2$, respectively. We decide that there is a crossing only when both $I_k(i), i = 1, \dots, n_k, k = 1, 2$ satisfy the conditions (7) and (8).

3 Crossing Length Estimation

3.1 Estimation of Camera Rotation Angle

In fact, the camera optical axis may be not exactly horizontal. In this subsection, we describe the method for estimating the camera rotation angle using the four consecutive feature points in a crossing image. We assume that the camera rotates with angle θ around the X_o -axis of the camera coordinate system (X_o, Y_o, Z_o) where the optical axis Z_o -axis and the X_o -axis are horizontal. It is assumed that the camera does not rotate around the optical axis. The θ is defined as the clockwise angle from the Z_o -axis. We also assume that the road surface is flat and horizontal. In this case, with rotated camera coordinate system, the road surface normal (i.e. the unit vector perpendicular to the road surface) \mathbf{n} is given by

$$\mathbf{n} = (n_x, n_y, n_z) = (0, \cos\theta, -\sin\theta). \quad (9)$$

We denote by (X, Y, Z) the rotated camera coordinate system. We use the four consecutive feature points extracted in a crossing image. We denote by y_i the image y -coordinate of the i -th feature point for the rotated camera coordinate system (X, Y, Z) where the image plane is given by $Z = f$, by Z_1 the horizontal distance from the camera center O to the first (nearest) feature point among the considered four consecutive points, by d the height of the camera center from the road surface and by Z the Z -coordinate of the first feature point in the rotated camera coordinate system (X, Y, Z) as illustrated in Fig.6. Then we have the following relation under the perspective projection:

$$Z = Z_1 \cos\theta + d \sin\theta, \quad (10)$$

$$y_i = f \frac{-(Z_1 + (i-1)b) \sin\theta + d \cos\theta}{(Z_1 + (i-1)b) \cos\theta + d \sin\theta}, \quad i = 1, \dots, 4. \quad (11)$$

From these relation, we can eliminate Z , b and d as follows:

$$1 = (g_2(\theta) - g_1(\theta)) / (g_4(\theta) - g_3(\theta)), \quad (12)$$

$$g_i(\theta) \equiv (f - y_i \tan\theta) / (f \tan\theta + y_i), \quad i = 1, \dots, 4. \quad (13)$$

For small θ , we can have an approximate relation in the following:

$$\begin{aligned} g_i(\theta) &= \frac{f}{y_i} \left(1 - \frac{y_i}{f} \tan\theta\right) / \left(1 + \frac{f}{y_i} \tan\theta\right) \\ &\simeq \frac{f}{y_i} \left(1 - \frac{y_i}{f} \tan\theta - \frac{f}{y_i} \tan\theta\right) \end{aligned} \quad (14)$$

Then we obtain the following approximate relations:

$$\tan\theta \simeq \frac{1}{f(\eta_4 + \eta_3)} \frac{\frac{\eta_2 - \eta_1}{\eta_4 - \eta_3} - 1}{\frac{\eta_2 - \eta_1}{\eta_4 - \eta_3} \frac{\eta_1 + \eta_2}{\eta_3 + \eta_4} - 1}, \quad (15)$$

$$\eta_i \equiv 1/y_i. \quad (16)$$

From this relation, we can estimate the angle θ .

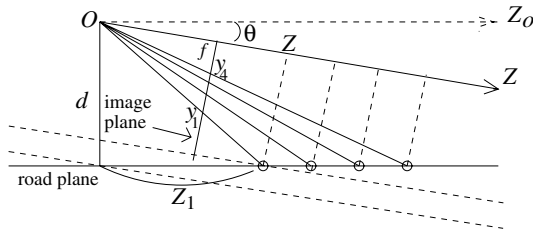


Fig. 6. The camera rotation angle θ

3.2 Extracting a Feature Point at the Half Crossing Length

In case of real pedestrian crossings, the road surface is not horizontal flat plane but has a gradient to drain away water due to rain. The road surface has the larger curvature at the farther point from the center of crossing corresponding to the half crossing length as illustrated in Fig 7. Hence at the half crossing length, the curvature of the road surface is considered to be minimal and the gradient of road surface is zero. We adopt the feature point with minimal curvature as the feature point at the half crossing length.

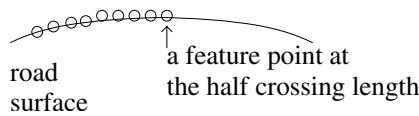


Fig. 7. The cross section of real road surface: the symbol “o” denotes a feature point

In case where there is a gradient of road surface, the θ estimated with equation (15) from four feature points on the road surface is the sum of the angle of camera θ_c and the angle of the road θ_r because the θ is the relative angle of the camera optical axis to the tangential plane of road surface where the four feature points exist as illustrated in Fig 8. It is assumed that the tangential plane of road surface is rotated by the angle θ_r around the X_o -axis. The angle of road θ_r is defined as the counter clockwise angle from the horizontal direction and the angle of camera θ_c is defined as the clockwise angle from the horizontal direction. Here, the angle θ is defined as the clockwise angle from the crossing direction in the tangential plane of the road surface as illustrated in Fig. 8. In case where the θ_c and θ_r are small, the derivative of $\tan \theta$ approximately coincides with the derivative of $\tan \theta_r$ because θ_c is constant with respect to position in road. The derivative of $\tan \theta$ with respect to length s on surface line in the crossing direction approximately coincides with the curvature $\frac{\Delta\theta_r}{\Delta s}$ of road surface in the crossing direction:

$$\frac{\Delta \tan \theta}{\Delta s} \simeq \frac{\Delta \tan \theta_r}{\Delta s} \simeq \frac{\Delta \theta_r}{\Delta s} \tag{17}$$

Hence the feature point at the half crossing length is obtained by finding the feature point with the minimum derivative of $\tan \theta$. The θ calculated with equation (15) from

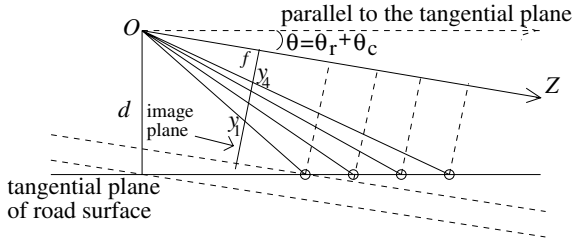


Fig. 8. The angle $\theta = \theta_r + \theta_c$ between the camera optical axis and the tangential plane of the road where the four feature points exist in case where there is a gradient of the road surface

four feature points at the half crossing length is the camera angle because the gradient of the road surface at the half crossing length is considered zero i.e. horizontal. The derivative of $\tan \theta$ is calculated from $\tan \theta$'s at adjacent feature points.

We adopt the more distant point among the two points corresponding to n_{1c} th and n_{2c} th feature points as the feature point at the half crossing length. Then we calculate the $\tan \theta$ of camera angle from the four consecutive feature points at the half crossing length and obtain the crossing length by the method described in the next subsection.

3.3 Crossing Length Estimation

Denote by y_{max} the y-coordinate of the feature point at the half crossing length in an observed image taken by a camera, which is rotated with an angle θ from the horizontal direction, and by y_1 the y-coordinate of the first (nearest) feature point in the image. Denote by n the number of white or black bands between y_1 and y_{max} and by h the horizontal distance to the half crossing length. Then we have the following relations:

$$y_1 = f \frac{-(h - nb)\sin\theta + d\cos\theta}{(h - nb)\cos\theta + d\sin\theta},$$

$$y_{max} = f \frac{-h\sin\theta + d\cos\theta}{h\cos\theta + d\sin\theta},$$

$$y_1 - y_{max} = f \frac{nb d(1 + \tan^2\theta)}{(h + d\tan\theta)(h + d\tan\theta - nb)},$$

$$h = 0.5 \left\{ nb(1 + \tan^2\theta) + \sqrt{(nb)^2(1 + \tan^2\theta)^2 + \frac{4nb(1 + \tan^2\theta)fd}{y_1 - y_{max}}} \right\} - d\tan\theta.$$

The crossing length ℓ of the present method is given by

$$\ell = 2 \times h.$$

It is assumed that an observer stands immediately before a pedestrian crossing. The distance without considering a camera angle was given by [11]. However, the present method considers the angle.

4 Experimental Results

To evaluate the performance of the proposed method for the detection of crossings and for measuring crossing length, we used 173 real images which include 100 images with crossings and 73 images without crossing, recorded by a digital camera. The digital camera was a Sony DSC-F707 with specifications of 1/3 in (8.47mm) CCD and 9.7mm

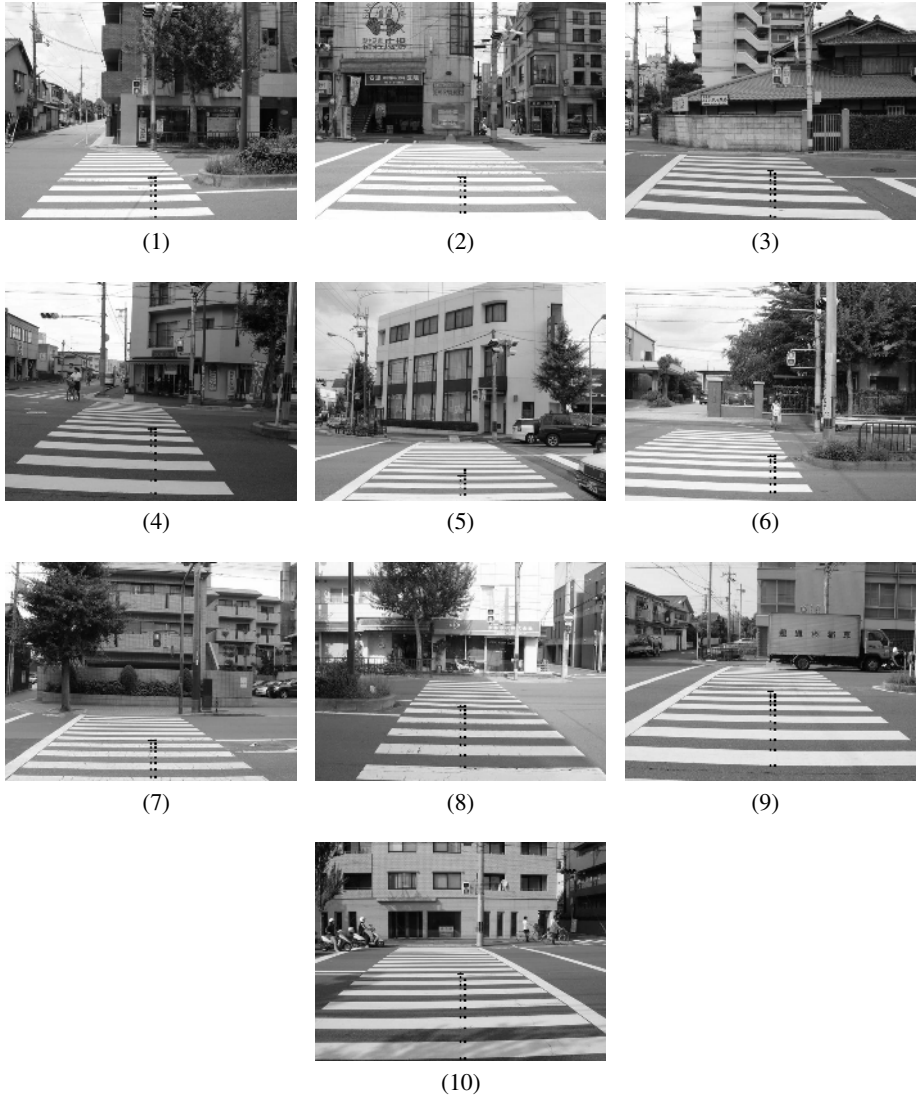


Fig. 9. Some images of frontal crossings with n_c extracted feature points. The n_c th point is the farthest point and is treated as the point at the half crossing length.

(724 pixels in case of 640×480 pixel image) focal length. For observation, the centre of the camera was set at the height of an observer's eye (1.65m) and images were obtained under various illumination conditions by observing in various weathers (except rain).

Some of images with crossings used for experiment are shown in Fig.9. In Fig.9, the extracted feature points are shown from the nearest point to the point at the half crossing length which is the n_c th feature point. The experimental results of $I_k, k = 1, 2$ defined by equation (5) are shown in Table 1 corresponding to Figs.9(1)~(10). In Table 1, the results satisfying relation (7) are shown in bold face. The proposed method is successful in detecting crossing for all 173 images.

Some of results for estimating the crossing length among the 100 images with crossings are shown in Table 2. The rms error of the estimated crossing lengths is 2.29m for 100 images with crossings. The maximal error is 6.58m. The reason for large errors in the estimated crossing length is that if the curvature of road surface is not symmetric or there is a local flat area which is different from area at the half crossing length then there occurs a local minimal curvature at a point different from the half crossing length and the feature point at the half crossing length is extracted at a different point from the true point as illustrated in Fig.9(3) which corresponds to the maximal error.

The cpu time of the proposed method in the present paper is 0.1s for detecting crossing and is 0.1s for measuring the crossing length by Pentium(R)M with 1.7GHz. On the other hand, by the method of [11], the cpu time is 1.3s for measuring the crossing length by Pentium M with 1.6GHz. In [11], the rms error is 2.28m for 32 images which are different from the 100 images with crossings in the present paper, but in the experiment

Table 1. Some of results for projective invariants, $I_1(i), i = 1, 2, \dots, n_1 - 3, I_2(i), i = 1, 2, \dots, n_2 - 3$, where n_1 and n_2 denote the numbers of extracted feature points on the two adjacent parallel vertical lines: Table numbers correspond to the numbers in Fig.9

i	I_1	I_2
1	0.237	0.237
2	0.265	0.265
3	0.249	0.249
4	0.235	0.235
5	0.241	0.241
$n_{1c}=8$ $n_{2c}=8$		

(1)

i	I_1	I_2
1	0.237	0.237
2	0.253	0.253
3	0.244	0.244
4	0.289	0.289
5	0.195	0.195
6	0.294	0.294
7	0.231	0.231
8	0.482	0.480
9	0.016	0.018
$n_{1c}=7$ $n_{2c}=7$		

(2)

i	I_1	I_2
1	0.253	0.286
2	0.234	0.212
3	0.267	0.274
4	0.248	0.233
5	0.248	0.277
6	0.247	0.250
7	0.246	0.214
8		0.498
$n_{1c}=10$ $n_{2c}=9$		

(3)

i	I_1	I_2
1	0.263	0.249
2	0.246	0.242
3	0.253	0.269
4	0.224	0.229
5	0.288	0.266
6	0.225	0.220
7	0.220	0.270
8	0.542	0.474
$n_{1c}=10$ $n_{2c}=10$		

(4)

i	I_1	I_2
1	0.220	0.235
2	0.307	0.259
3	0.206	0.227
4	0.260	0.308
5	0.491	0.412
6	0.024	0.010
7	0.348	0.705
8	0.234	0.111
9		0.365
$n_{1c}=5$ $n_{2c}=7$		

(5)

i	I_1	I_2
1	0.247	0.264
2	0.233	0.226
3	0.268	0.283
4	0.230	0.222
5	0.281	0.238
6	0.446	0.521
7	0.017	0.011
8	0.675	0.725
9		0.020
$n_{1c}=8$ $n_{2c}=8$		

(6)

i	I_1	I_2
1	0.236	0.236
2	0.278	0.278
3	0.214	0.214
4	0.286	0.286
5	0.219	0.219
6	0.525	0.513
7		0.022
$n_{1c}=8$ $n_{2c}=8$		

(7)

i	I_1	I_2
1	0.240	0.244
2	0.274	0.259
3	0.240	0.247
4	0.235	0.235
5	0.280	0.267
6	0.211	0.248
7	0.270	0.230
8		0.509
$n_{1c}=10$ $n_{2c}=10$		

(8)

i	I_1	I_2
1	0.270	0.270
2	0.239	0.231
3	0.250	0.274
4	0.253	0.229
5	0.258	0.266
6	0.235	0.235
7	0.259	0.259
8	0.247	0.247
9	0.246	
$n_{1c}=12$ $n_{2c}=11$		

(9)

i	I_1	I_2
1	0.242	0.254
2	0.253	0.249
3	0.242	0.242
4	0.249	0.249
5	0.265	0.265
6	0.236	0.236
7	0.258	0.273
8	0.248	0.206
9	0.247	0.294
$n_{1c}=12$ $n_{2c}=10$		

(10)

Table 2. Some of results for crossing length estimation

no	true length[m]	estimate[m]	error[m]
716 - 1	14.22	15.92	1.70
716 - 2	14.22	16.94	2.72
716 - 3	16.92	18.70	1.78
716 - 4	19.57	18.47	-1.10
716 - 5	14.69	16.63	1.94
716 - 6	18.36	17.89	-0.47
716 - 7	18.36	17.96	-0.40
716 - 8	22.40	17.44	-4.96
716 - 9	17.68	18.13	0.45
716 - 10	17.70	15.76	-1.94
716 - 11	19.57	15.64	-3.93
716 - 12	14.18	17.97	3.79
716 - 13	14.20	16.66	2.46
716 - 14	16.05	18.74	2.69
716 - 15	11.93	18.51	6.58
716 - 16	15.23	15.84	0.61
716 - 17	16.92	17.95	1.03
716 - 18	14.26	16.21	1.95
716 - 19	16.71	17.77	1.06
716 - 20	15.28	16.23	0.95
716 - 21	17.55	17.30	-0.25
716 - 22	17.70	18.48	0.78
716 - 23	9.73	15.97	6.24
716 - 24	14.22	16.07	1.85
716 - 25	20.69	18.55	-2.14

no	true length[m]	estimate[m]	error[m]
716 - 26	20.41	18.61	-1.80
716 - 27	20.47	19.29	-1.18
716 - 28	14.23	17.45	3.22
716 - 29	12.75	17.16	4.41
716 - 30	14.78	17.29	2.51
716 - 31	21.52	18.29	-3.23
716 - 32	14.22	17.70	3.48
716 - 33	14.01	16.73	2.72
716 - 34	14.00	14.55	0.55
716 - 35	11.17	17.16	5.99
716 - 36	14.32	17.05	2.73
716 - 37	14.08	15.53	1.45
716 - 38	14.23	17.94	3.71
716 - 39	18.87	18.36	-0.51
716 - 40	17.80	18.57	0.77
716 - 41	19.55	17.82	-1.73
716 - 42	17.50	18.22	0.72
716 - 43	15.05	16.61	1.56
716 - 44	14.75	15.92	1.17
716 - 45	14.80	17.11	2.31
716 - 46	21.00	18.55	-2.45
716 - 47	19.20	16.30	-2.90
716 - 48	17.73	17.86	0.13
716 - 49	17.72	17.77	0.05
716 - 50	17.45	17.14	-0.31

for the 100 images in the present paper the method of [11] fails to extract necessary features and can not estimate the crossing length for one image among the 100 images.

5 Conclusion

We have proposed a method for detecting frontal pedestrian crossings and estimating its length from image data obtained with a single camera as a travel aid for the blind. The method has been successful in detecting a pedestrian crossing for all 173 real images which include 100 images with crossings. The rms error in estimating the crossing length is 2.29m for the 100 images with crossings.

Acknowledgments

The authors are grateful for the support of Japan Society for the Promotion of Science under Grants-in-Aid for Scientific Research (No.16500110 and No.03232).

References

1. Kay L 1974 *A sonar aid to enhance spatial perception of the blind, engineering design and evaluation*, Radio and Electron. Eng. **44** 605-29
2. Morrisette D L et al 1981 *A follow-up study of the mowat sensor's applications, frequency of use, and maintenance reliability*, J. Vis. Impairment and Blindness **75** 244-7
3. Benjamin J M 1973 *The new C-5 laser cane for the blind*, Proc. Carnahan Conf. on Electronic Prosthetics 77-82
4. Shoval S, Borenstein J and Koren Y 1998 *Auditory guidance with the navbelt-a computerized travel aid for the blind*, IEEE Trans. Syst. Man Cybern. C **28** 459-67
5. Tajima T, Aotani T, Kurauchi K and Ohkubo H 2002 *Evaluation of pedestrian information and communication systems-A for visually impaired persons*, Proc. 17th CSUN conf. 1-7.

6. Meijar P B L *Vision technology for the totally blind*, [Online] Available: <http://www.seeingwithsound.com/>.
7. Stephen Se 2000 *Zebra-crossing detection for the partially sighted*, Proc.IEEE Computer Society Conf. on Computer Vision and Pattern Recognition (CVPR), (South Carolina, USA, Jun. 2000), **2** 211-217
8. Shioyama T, Wu H, Nakamura N and Kitawaki S 2002 *Measurement of the length of pedestrian crossings and detection of traffic lights from image data*, Meas. Sci. Technol. **13** 1450-7
9. Shioyama T and Uddin M S 2004 *Detection of pedestrian crossings with projective invariants from image data* Meas. Sci. Technol. **15** 2400-5
10. Weiss I 1993 *Geometric invariants and object recognition*, Int. J. of Comp. Vision **10** 207-31
11. Uddin M S and Shioyama T 2004 *Measurement of the length of pedestrian crossings from image data* Proc. Int. Symp. on Photo-Optical Instrumentation Engineering(SPIE2004,Photonics North, Ottawa) **5578** 498-508

A Novel Stochastic Attributed Relational Graph Matching Based on Relation Vector Space Analysis

Bo Gun Park, Kyoung Mu Lee, and Sang Uk Lee

School of Electrical Eng., ASRI, Seoul National University,
151-600, Seoul, Korea

gun@diehard.snu.ac.kr, kyoungmu@snu.ac.kr, sanguk@sting.snu.ac.kr

Abstract. In this paper, we propose a novel stochastic attributed relational graph (SARG) matching algorithm in order to cope with possible distortions due to noise and occlusion. The support flow and the correspondence measure between nodes are defined and estimated by analyzing the distribution of the attribute vectors in the relation vector space. And then the candidate subgraphs are extracted and ordered according to the correspondence measure. Missing nodes for each candidates are identified by the iterative voting scheme through an error analysis, and then the final subgraph matching is carried out effectively by excluding them. Experimental results on the synthetic ARGs demonstrate that the proposed SARG matching algorithm is quite robust and efficient even in the noisy environment. Comparative evaluation results also show that it gives superior performance compared to other conventional graph matching approaches.

1 Introduction

Object recognition has been one of the most challenging problems in computer vision for several decades. Since an object can be distinguished from other objects by its own structure, the description of a structured object in terms of its primitives parts and mutual relations between them, has been an important issue in artificial intelligence. Due to its representational power, the graph models has been widely used for formulating such structured abstract pattern. Among various kinds of graphs, attributed relational graph (ARG) has been considered as the most effective data structures, which consists of nodes and attribute vectors for encoding unary properties and mutual relations [6] [9].

So far, a great deal of works have been reported in the literature for developing efficient and robust graph matching techniques. According to [8], the graph matching algorithms can be categorized into two classes: Search-based approach and optimization-based approach. Search-based approaches [6][7][9][10] construct the states-spaces representing graph matching status, which are searched with techniques similar to the tree search [6] in order to find the optimal solution satisfying some criteria. In general, most of them have the

exponential complexity in the worst case, although a few methods [9] showed a high-order polynomial complexity with the help of heuristics. On the other hand, optimization-based approaches including relaxation labeling [4] [5], simulated annealing [11], genetic algorithms [12], and interpolator-based algorithms [15] [16] consider the graph matching as one of the energy minimization problems. Most of them try to find the sub-optimal solution in a continuous state space based on heuristic methods, that usually take polynomial time [4] [5]. In real applications, however, the graph representation that exhibits the morphology of the graph and the attributes is often corrupted by noise or imprecise abstraction. In order to cope with these uncertainties in the graph structure and attributes, probabilistic graph models have been introduced [1] [4] [5]. Wong *et al.* [1] interpreted an ensemble of ARGs as the outcomes of a random graph, in which nodes, edges, and all attributes were random variables with the probability density function trained from sample graphs by using supervised learning. Sanfeliu *et al.* [2] [3] presented the function-described graph (FDG) as the extension of the random graph, which is the compact representation of a set of ARGs. In order to alleviate the statistical independence of nodes and edges, some qualitative knowledge of the second-order probabilities of the elements was incorporated into FDGs. Recently, a new partial ARG matching approach was proposed that introduced the relation vector space concept to cope with large variation of attributes and partial matching [18]. However, since it can not incorporate the unary information into the model, its performance and the applications are very limited.

In this paper, we propose a new robust stochastic partial ARG matching technique which utilizes both the binary relation and unary attribute information in the relational vector space. The proposed graph matching algorithm consists of two phases: In the first stage, the candidate subgraphs are extracted and sorted according to the correspondence measure, which are based on the stochastic analysis in the relation vector space. This process significantly reduces the number of possible matches, and in the result the proposed algorithm has the polynomial computational complexity. Then, missing nodes for each candidates are identified by the iterative voting scheme through the error analysis until no more node is found to be missed, and the final subgraph matching is carried out effectively by excluding them.

2 Attributed Relational Graph

2.1 Definition of ARG

Let us define an ARG with N nodes as

$$\begin{aligned} \mathcal{G} &= (\mathcal{V}, \mathcal{E}, \mathcal{U}, \mathcal{B}, \mathcal{F}), & (1) \\ \mathcal{V} &= \{v_1, \dots, v_N\}, \quad \mathcal{E} = \{e_{ij} | i = 1, \dots, N, j = 1, \dots, N, i \neq j\}, \\ \mathcal{U} &= \{\mathbf{a}_i | i = 1, \dots, N\}, \quad \mathcal{B} = \{\mathbf{r}_{ij} | i = 1, \dots, N, j = 1, \dots, N, i \neq j\}, \\ \mathcal{F} &= \{\mathcal{R}_i | \mathcal{R}_i = \{\mathbf{r}_{ij} | v_i, v_j \in \mathcal{V}, i \neq j\}, i = 1, \dots, N\}, \end{aligned}$$

where \mathcal{V} and \mathcal{E} are the sets of nodes and edges in the graph, respectively. If the edge between node v_i and v_j exists, e_{ij} is equal to 1, otherwise, it is 0. And \mathbf{a}_i is an N_U -dimensional unary attribute vector of the node v_i , and \mathbf{r}_{ij} denotes the N_B -dimensional binary attribute vector of the edge connecting node v_i and v_j . \mathcal{F} is the set of relation vector spaces that encode the structural information centered at each node [18].

2.2 Preliminaries of ARG Matching

Definition 1 (Attribute matrices). *The unary attribute matrices $\mathbf{U}_i \in R^{N \times 1}$ and the binary attribute matrices $\mathbf{B}_j \in R^{N \times N}$ of a graph \mathcal{G} with N nodes are defined by*

$$\mathbf{U}_i = [\mathbf{a}_1(i) \cdots \mathbf{a}_N(i)]^T, i = 1, \dots, N_U, \tag{2}$$

$$\mathbf{B}_i = \begin{bmatrix} \mathbf{r}_{11}(i) & \cdots & \mathbf{r}_{1N}(i) \\ \vdots & \mathbf{r}_{lm}(i) & \vdots \\ \mathbf{r}_{N1}(i) & \cdots & \mathbf{r}_{NN}(i) \end{bmatrix}, i = 1, \dots, N_B, \tag{3}$$

where $\mathbf{a}_i(k)$ and $\mathbf{r}_{ij}(k)$ represent the k -th elements of \mathbf{a}_i and \mathbf{r}_{ij} , respectively.

Assume that two ARGs, \mathcal{G} and \mathcal{G}' , are given. Messmer *et al.* [7] rigorously defined the graph isomorphism and the subgraph isomorphism by establishing the linear relation between the attribute matrices of two graphs through the permutation matrix. However, they only considered the ideal case, that is, without noise. Thus, in this section, we generalize the graph isomorphism and the subgraph isomorphism to cope with the corruption due to noise.

Definition 2 (Graph isomorphism). *Two graphs \mathcal{G}' and \mathcal{G} are called isomorphic if there exists an $N' \times N$ permutation matrix \mathbf{P} such that*

$$\begin{aligned} \mathbf{U}'_i &= \mathbf{P}\mathbf{U}_i + \epsilon \mathbf{C}_i^U \mathbf{N}_i^U, \\ \mathbf{B}'_i &= \mathbf{P}\mathbf{B}_i\mathbf{P}^T + \epsilon \begin{bmatrix} \cdots & & \\ & Sum(\mathbf{C}_{lm}^B * \mathbf{N}_i^B) & \\ & & \cdots \end{bmatrix}, \end{aligned} \tag{4}$$

where $\mathbf{N}_i^U \in R^{N' \times 1}$ and $\mathbf{N}_j^B \in R^{N' \times N'}$ are the noise matrices of which components are statistically independent. $\mathbf{C}_i^U \in R^{N' \times N'}$ and $\mathbf{C}_{lm}^B \in R^{N' \times N'}$ are the noise correlation matrices.

In (4), operator $'*$ ' represents the component-wise multiplication operation, and the function $'Sum(\cdot)'$ is the sum of all elements of the matrix \cdot .

Definition 3 (Subgraph isomorphism). *Two graphs \mathcal{G}' and \mathcal{G} are called subgraph isomorphic if there exist two sub-graphs, $\hat{\mathcal{G}} = (\hat{\mathcal{V}}, \hat{\mathcal{E}}, \hat{\mathcal{U}}, \hat{\mathcal{B}}, \hat{\mathcal{F}}) \subset \mathcal{G}$ and $\hat{\mathcal{G}}' = (\hat{\mathcal{V}}', \hat{\mathcal{E}}', \hat{\mathcal{U}}', \hat{\mathcal{B}}', \hat{\mathcal{F}}') \subset \mathcal{G}'$ that are isomorphic.*

From the above definitions, the ARG matching problem can be thought as the ARG matching by the subgraph isomorphism, that is the extraction of subgraphs and the inference of \mathbf{P} satisfying (4). In other words, it can be termed as “correspondence problem” between two ARGs [10] [16].

3 Proposed SARG Matching Algorithm

3.1 Correspondence Measure

As stated in Section 2, the graph matching problem can be transformed into the correspondence problem. So, let us define some basic concepts related to the correspondence. Assume that a reference graph $\mathcal{G}_M = (\mathcal{V}^{\mathcal{G}_M}, \mathcal{E}^{\mathcal{G}_M}, \mathcal{U}^{\mathcal{G}_M}, \mathcal{B}^{\mathcal{G}_M}, \mathcal{F}^{\mathcal{G}_M})$ and an input graph $\mathcal{G}_I = (\mathcal{V}^{\mathcal{G}_I}, \mathcal{E}^{\mathcal{G}_I}, \mathcal{U}^{\mathcal{G}_I}, \mathcal{B}^{\mathcal{G}_I}, \mathcal{F}^{\mathcal{G}_I})$ are given.

Definition 4 (Correspondence). *If i -th node of \mathcal{G}_M and the l -th node of \mathcal{G}_I match by each other, then there exist a correspondence between them, and it is denoted by*

$$v_i^{\mathcal{G}_M} \leftrightarrow v_l^{\mathcal{G}_I} \text{ or } v_l^{\mathcal{G}_I} = Cor(v_i^{\mathcal{G}_M}) \text{ or } l = Cor(v_i^{\mathcal{G}_M}). \tag{5}$$

Definition 5 (Stochastic neighborhood). *The stochastic neighborhood of binary attribute vector $\mathbf{r}_{ij}^{\mathcal{G}_M}$ in the relation vector space $\mathcal{R}_l^{\mathcal{G}_I}$ under the assumption of $v_i^{\mathcal{G}_M} \leftrightarrow v_l^{\mathcal{G}_I}$ is defined by*

$$\begin{aligned} \mathcal{N}_{\mathcal{R}_l^{\mathcal{G}_I}}(\mathbf{r}_{ij}^{\mathcal{G}_M}) &= \{v_m^{\mathcal{G}_I} \mid \|\mathbf{r}_{lm}^{\mathcal{G}_I} - \mathbf{r}_{ij}^{\mathcal{G}_M}\|_{prob} < \Delta\}, \\ &= \{v_m^{\mathcal{G}_I} \mid p(\mathbf{r}_{lm}^{\mathcal{G}_I} - \mathbf{r}_{ij}^{\mathcal{G}_M}) \times p(\mathbf{a}_m^{\mathcal{G}_I} - \mathbf{a}_j^{\mathcal{G}_M}) < \Delta\}. \end{aligned} \tag{6}$$

Definition 6 (Support flow). *The support flow from $v_j^{\mathcal{G}_M}$ to $v_i^{\mathcal{G}_M}$ under the assumption of $v_i^{\mathcal{G}_M} \leftrightarrow v_l^{\mathcal{G}_I}$ is defined by*

$$\begin{aligned} \mathcal{F}_{sup}(v_j^{\mathcal{G}_M} \mid v_i^{\mathcal{G}_M}, v_l^{\mathcal{G}_I}) &= \\ \left\{ \begin{array}{ll} \sum_{\substack{v_m^{\mathcal{G}_I} \in \\ \mathcal{N}_{\mathcal{R}_l^{\mathcal{G}_I}}(\mathbf{r}_{ij}^{\mathcal{G}_M})}} \frac{p(\mathbf{r}_{lm}^{\mathcal{G}_I} - \mathbf{r}_{ij}^{\mathcal{G}_M}) \cdot \mathcal{M}_{cor}(v_j^{\mathcal{G}_M}, v_m^{\mathcal{G}_I})}{|\mathcal{N}_{\mathcal{R}_l^{\mathcal{G}_I}}(\mathbf{r}_{ij}^{\mathcal{G}_M})|} & \text{if } \mathcal{N}_{\mathcal{R}_l^{\mathcal{G}_I}}(\mathbf{r}_{ij}^{\mathcal{G}_M}) \neq \emptyset, \\ \text{Max } p(\mathbf{r}_{lm}^{\mathcal{G}_I} - \mathbf{r}_{ij}^{\mathcal{G}_M}) \cdot \mathcal{M}_{cor}(v_j^{\mathcal{G}_M}, v_m^{\mathcal{G}_I}) & \text{if } \mathcal{N}_{\mathcal{R}_l^{\mathcal{G}_I}}(\mathbf{r}_{ij}^{\mathcal{G}_M}) = \emptyset, \end{array} \right. \tag{7} \end{aligned}$$

where $\mathcal{M}_{cor}(v_j^{\mathcal{G}_M}, v_m^{\mathcal{G}_I})$ is a correspondence measure between $v_j^{\mathcal{G}_M}$ and $v_m^{\mathcal{G}_I}$

The support flow in (7) represents how much a neighboring node supports the given node correspondence. Fig. 1 shows an example of the stochastic neighborhoods and support flows. Provided that i -th node of \mathcal{G}_M corresponds to the l -th node of \mathcal{G}_I , $\mathcal{R}_l^{\mathcal{G}_I}$ must be similar to $\mathcal{R}_i^{\mathcal{G}_M}$. From the definition in (7), it is noted that as the similarity between two relation vector spaces of the corresponding nodes increases, the sum of the support flows from other nodes also increases. As a result, the sum of the support flows can be used as an indication for the correspondence between two nodes. Based on this observation, the *correspondence measure* is defined as the average of the support flows from other nodes.

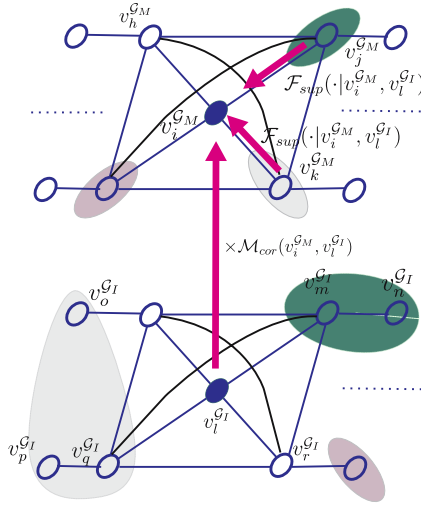


Fig. 1. Graphical representation of stochastic neighborhoods and support flows. Corresponding neighborhoods are represented by the colored regions, and the arrows show the support flows from neighbor nodes to the specific node in a reference graph.

Definition 7 (Correspondence measure). *The correspondence measure between $v_i^{G_M}$ and $v_l^{G_I}$ is defined by*

$$\mathcal{M}_{cor}(v_i^{G_M}, v_l^{G_I}) = \frac{\sum_{j=1, j \neq i}^{N^{G_M}} \mathcal{F}_{sup}(v_j^{G_M} | v_i^{G_M}, v_l^{G_I}) \cdot \mathcal{M}_{cor}(v_i^{G_M}, v_l^{G_I})}{(N^{G_M} - 1)}. \tag{8}$$

Note that in order to embed the structural consistency, the actual correspondence measure is calculated iteratively through an updating process using (7) and (8) as follows.

1) *Initialize:* $k = 0$

$$\mathcal{M}_{cor}^{(0)}(v_i^{G_M}, v_l^{G_I}) = p(\mathbf{a}_l^{G_I} - \mathbf{a}_i^{G_M}). \tag{9}$$

2) *Updating process:* $k \geq 1$

$$\mathcal{F}_{sup}^{(k)}(v_j^{G_M} | v_i^{G_M}, v_l^{G_I}) = \begin{cases} \sum_{\substack{v_m^{G_I} \in \\ \mathcal{N}_{\mathcal{R}_l^{G_I}}(\mathbf{r}_{ij}^{G_M})}} \frac{p(\mathbf{r}_{lm}^{G_I} - \mathbf{r}_{ij}^{G_M}) \cdot \mathcal{M}_{cor}^{(k-1)}(v_j^{G_M}, v_m^{G_I})}{|\mathcal{N}_{\mathcal{R}_l^{G_I}}(\mathbf{r}_{ij}^{G_M})|} & \text{if } \mathcal{N}_{\mathcal{R}_l^{G_I}}(\mathbf{r}_{ij}^{G_M}) \neq \emptyset, \\ \text{Max } p(\mathbf{r}_{lm}^{G_I} - \mathbf{r}_{ij}^{G_M}) \cdot \mathcal{M}_{cor}^{(k-1)}(v_j^{G_M}, v_m^{G_I}) & \text{if } \mathcal{N}_{\mathcal{R}_l^{G_I}}(\mathbf{r}_{ij}^{G_M}) = \emptyset, \end{cases} \tag{10}$$

$$\mathcal{M}_{cor}^{(k)}(v_i^{G_M}, v_l^{G_I}) = \frac{\sum_{j=1, j \neq i}^{N^{G_M}} \mathcal{F}_{sup}^{(k)}(v_j^{G_M} | v_i^{G_M}, v_l^{G_I}) \cdot \mathcal{M}_{cor}^{(k-1)}(v_i^{G_M}, v_l^{G_I})}{(N^{G_M} - 1)}. \tag{11}$$

3.2 Selection of Candidate Sub-graphs

In the combinatorial graph matching methods, constructing meaningful candidate subgraphs and ordering them are important issues. The proposed SARG matching algorithm extracts and sorts candidate subgraphs by measuring correspondence stochastically in the relation vector space.

Provided that i -th node of \mathcal{G}_M corresponds to the l -th node of \mathcal{G}_I , the subspace of $\mathcal{R}_l^{\mathcal{G}_I}$ that is similar to $\mathcal{R}_i^{\mathcal{G}_M}$ can be constructed as

$$\hat{\mathcal{R}}_l^{\mathcal{G}_I} = \{\mathbf{r}_{lm}^{\mathcal{G}_I} | v_m^{\mathcal{G}_I} \in \mathcal{N}_{\mathcal{R}_l^{\mathcal{G}_I}}(\mathbf{r}_{ij}^{\mathcal{G}_M}), j = 1, \dots, N^{\mathcal{G}_M}, j \neq i\}. \tag{12}$$

Then, a set of initial candidate subgraphs subject to $v_i^{\mathcal{G}_M} \leftrightarrow v_l^{\mathcal{G}_I}$ can be constructed by selecting one node in each neighbor $\mathcal{N}_{\mathcal{R}_l^{\mathcal{G}_I}}(\mathbf{r}_{ij}^{\mathcal{G}_M})$ as

$$\begin{aligned} \mathcal{A}(v_i^{\mathcal{G}_M}, v_l^{\mathcal{G}_I}) = & \{(v_{k_1}^{\mathcal{G}_I}, \dots, v_{k_{i-1}}^{\mathcal{G}_I}, v_l^{\mathcal{G}_I}, v_{k_{i+1}}^{\mathcal{G}_I}, \dots, v_{k_N^{\mathcal{G}_M}}^{\mathcal{G}_I}) | \\ & v_{k_j}^{\mathcal{G}_I} \in \mathcal{N}_{\mathcal{R}_l^{\mathcal{G}_I}}(\mathbf{r}_{ij}^{\mathcal{G}_M}), j = 1, 2, \dots, N^{\mathcal{G}_M}, j \neq i\}. \end{aligned} \tag{13}$$

Once all $\mathcal{A}(v_i^{\mathcal{G}_M}, v_l^{\mathcal{G}_I})$, for $i = 1, \dots, N^{\mathcal{G}_M}$, and $l = 1, \dots, N^{\mathcal{G}_I}$, are obtained, the total initial candidate subgraphs are given by the union of all sets $\mathcal{A}(v_i^{\mathcal{G}_M}, v_l^{\mathcal{G}_I})$;

$$\mathcal{A}_T = \bigcup_{i,l} \mathcal{A}(v_i^{\mathcal{G}_M}, v_l^{\mathcal{G}_I}). \tag{14}$$

Now, meaningless initial subgraphs can be excluded out from (14) by ordering them according to the correspondence measure defined in (8), and the final candidate subgraphs with high priority are given by

$$\hat{\mathcal{A}}_T = \{\mathcal{A}(v_i^{\mathcal{G}_M}, v_l^{\mathcal{G}_I}) | \mathcal{M}_{cor}(v_i^{\mathcal{G}_M}, v_l^{\mathcal{G}_I}) \geq \alpha\}. \tag{15}$$

3.3 Missing Node Detection and Correction

Assume that a candidate subgraph $\mathcal{G}_C \in \hat{\mathcal{A}}_T$ and a reference graph \mathcal{G}_M are given. Each node in \mathcal{G}_C has one-to-one correspondence to a node with same index in \mathcal{G}_M , and this relation is denoted by $v_i^{\mathcal{G}_C} \leftrightarrow v_i^{\mathcal{G}_M}$ or $v_i^{\mathcal{G}_C} = Cor(v_i^{\mathcal{G}_M})$.

Then the missing nodes can be detected by constructing and analyzing the node loss vector given by,

$$\mathcal{L} = [l(1) \dots l(N^{\mathcal{G}_M})]^T, \tag{16}$$

where $l(i)$ is the number of nodes that satisfies the inequality $p(\mathbf{r}_{ij}^{\mathcal{G}_C} - \mathbf{r}_{ij}^{\mathcal{G}_M}) < p_{thres}$. Each detected missing node has the correspondence to NULL node ($v_i^{\mathcal{G}_M} \leftrightarrow v_0$).

For example, after detecting missing nodes from one of the candidate subgraphs in Fig. 2 (a), $\mathcal{V}^{\mathcal{G}_M}$ is partitioned as two sets as $\mathcal{V}^{\mathcal{G}_M}$ and $\mathcal{V}_0^{\mathcal{G}_M} = \{v_1^{\mathcal{G}_M}, v_2^{\mathcal{G}_M}\}$ as shown in Fig. 2 (b), where dark circles connected by the arrow represent the corresponding node pair and the dotted circles mean the detected missing nodes. However, it is certain that $v_1^{\mathcal{G}_M}$ has the correspondence to one node in \mathcal{G}_I instead

of the NULL node. Thus, we propose the additional procedure to correct wrong correspondence to the NULL node as $v_1^{\mathcal{G}_M}$.

Now, the node set of \mathcal{G}_M , $\mathcal{V}^{\mathcal{G}_M}$ can be partitioned into two sets as

$$\mathcal{V}^{\mathcal{G}_M} = \mathcal{V}_C^{\mathcal{G}_M} + \mathcal{V}_0^{\mathcal{G}_M}, \tag{17}$$

where all nodes in $\mathcal{V}_0^{\mathcal{G}_M}$ have the correspondence to the NULL node, v_0 . In order to reduce the false detection and make as many correspondences as possible, we recompute the correspondence measures for the nodes in $\mathcal{V}_0^{\mathcal{G}_M}$ again by using

$$\begin{aligned} \mathcal{M}_{cor}(v_i^{\mathcal{G}_M}, v_l^{\mathcal{G}_I}) &= \sum_{v_j^{\mathcal{G}_M} \in \mathcal{V}_C^{\mathcal{G}_M}, j \neq i} p(\mathbf{r}_{ij}^{\mathcal{G}_M} - \mathbf{r}_{lCor(v_j^{\mathcal{G}_M})}^{\mathcal{G}_M}) \\ &\cdot p(\mathbf{a}_i^{\mathcal{G}_M} - \mathbf{a}_l^{\mathcal{G}_I}) \cdot p(\mathbf{a}_j^{\mathcal{G}_M} - \mathbf{a}_{Cor(v_j^{\mathcal{G}_M})}^{\mathcal{G}_I}), \text{ for } v_i^{\mathcal{G}_M} \in \mathcal{V}_0^{\mathcal{G}_M}. \end{aligned} \tag{18}$$

Then, we determine the final correspondence for each node in $\mathcal{V}_0^{\mathcal{G}_M}$ by

$$\begin{aligned} v_i^{\mathcal{G}_M} &\leftrightarrow \\ \left\{ \begin{array}{l} v_l^{\mathcal{G}_I} \quad \text{if } v_l^{\mathcal{G}_I} = \arg \max_{v_k^{\mathcal{G}_I}} \mathcal{M}_{cor}(v_i^{\mathcal{G}_M}, v_k^{\mathcal{G}_I}) \text{ and } \mathcal{M}_{cor}(v_i^{\mathcal{G}_M}, v_l^{\mathcal{G}_I}) \geq \beta, \\ v_0 \quad \text{otherwise.} \end{array} \right. \end{aligned} \tag{19}$$

Fig. 2 shows an example. By recomputing the correspondence measure based on the strong correspondences as described above, missing nodes are detected and the final correct correspondences can be obtained as in Fig. 2 (c).

3.4 Matching

Once all the candidate subgraphs are corrected, we can find the subgraph that best matches the model using the similarity measure given by

$$\begin{aligned} S(\mathcal{G}_M, \mathcal{G}_C) &= \sum_{i=1}^{N^{\mathcal{G}_M}} \mathcal{D}(\mathcal{R}_i) \cdot p(\mathbf{a}_i^{\mathcal{G}_1} - \mathbf{a}_i^{\mathcal{G}_2}) \cdot \omega_i \\ &= \sum_{i=1}^{N^{\mathcal{G}_M}} \omega_i \cdot p(\mathbf{a}_i^{\mathcal{G}_1} - \mathbf{a}_i^{\mathcal{G}_2}) \cdot \left[\prod_{j=1, j \neq i}^{N^{\mathcal{G}_M}} p(\mathbf{r}_{ij}^{\mathcal{G}_C} - \mathbf{r}_{ij}^{\mathcal{G}_M}) \cdot \gamma_{ij} \right], \end{aligned} \tag{20}$$

where $\mathcal{D}(\mathcal{R}_i)$ is a function to measure the difference between the relation vector space $\mathcal{R}_i^{\mathcal{G}_1}$ of \mathcal{G}_1 and $\mathcal{R}_i^{\mathcal{G}_2}$ of \mathcal{G}_2 , and ω_i and γ_{ij} are the weighting for the i -th node and the binary relation between node v_i and v_j , respectively.

Then, the best matched subgraph is selected as follows.

$$\text{Matched Graph} = \arg \max_{\mathcal{G}_C \in \hat{\mathcal{A}}_T} S(\mathcal{G}_M, \mathcal{G}_C). \tag{21}$$

4 Computational Complexity

In this section, the computational complexity of the proposed SARG matching algorithm is analyzed. The computation of the proposed SARG matching

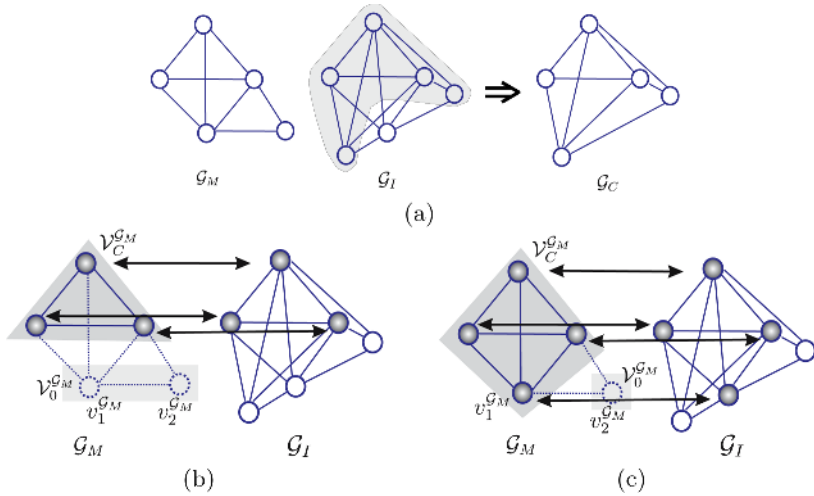


Fig. 2. Missing node detection and correcting process. (a) One of the candidate subgraphs are selected and (b) missing nodes are detected by analyzing the node loss vector, where the corresponding node pair is connected by the arrow and the dotted circles mean missing nodes. By recomputing the correspondence measure based on the strong correspondence, (c) the final correspondences are obtained.

algorithm consists of two parts: (1) Selecting the candidate subgraphs and (2) identifying missing node and matching. In the first stage, the stochastic neighborhoods should be constructed before computing the correspondence measures, and it needs iterative calculation. Then the candidate subgraphs are extracted and sorted according to the correspondence measure. Assume that a reference graph \mathcal{G}_M with $N^{\mathcal{G}_M}$ nodes and an input graph \mathcal{G}_I with $N^{\mathcal{G}_I}$ nodes are given, and K iterations are performed to update the correspondence measure. Construction of the stochastic neighborhoods for all node pairs is proportional to the square of the total possible number of correspondences, $N^{\mathcal{G}_M} N^{\mathcal{G}_I}$, that is approximately equal to $O(N^{\mathcal{G}_M^2} N^{\mathcal{G}_I^2})$. And since, at each iteration step, $N^{\mathcal{G}_M}$ support flows in (7) should be evaluated per each correspondence pair, the cost of computing the correspondence measures becomes $O(K \cdot N^{\mathcal{G}_M^2} N^{\mathcal{G}_I})$. And, the total cost for extracting and sorting the candidate subgraphs is proportional to the total number of correspondences, i.e., $O(N^{\mathcal{G}_M} N^{\mathcal{G}_I})$. In the result, the computational complexity of the first part is

$$\begin{aligned}
 T_1 &= T_{neighbor} + T_{correspondence} + T_{candidate} \\
 &= O(N^{\mathcal{G}_M^2} N^{\mathcal{G}_I^2}) + O(K \cdot N^{\mathcal{G}_M^2} N^{\mathcal{G}_I}) + O(N^{\mathcal{G}_M} N^{\mathcal{G}_I}) = O(N^{\mathcal{G}_M^2} N^{\mathcal{G}_I^2}).
 \end{aligned}$$

In the second stage, for each candidate subgraph, the missing node detection and correction processes are carried out first, and then matching is done by evaluating the similarity between two ARGs. Therefore the computational complexity of the second part is proportional to the product of the number of candidate subgraphs and the computational cost required for one candidate subgraph. Note that for a given candidate subgraph, the costs for detecting missing nodes

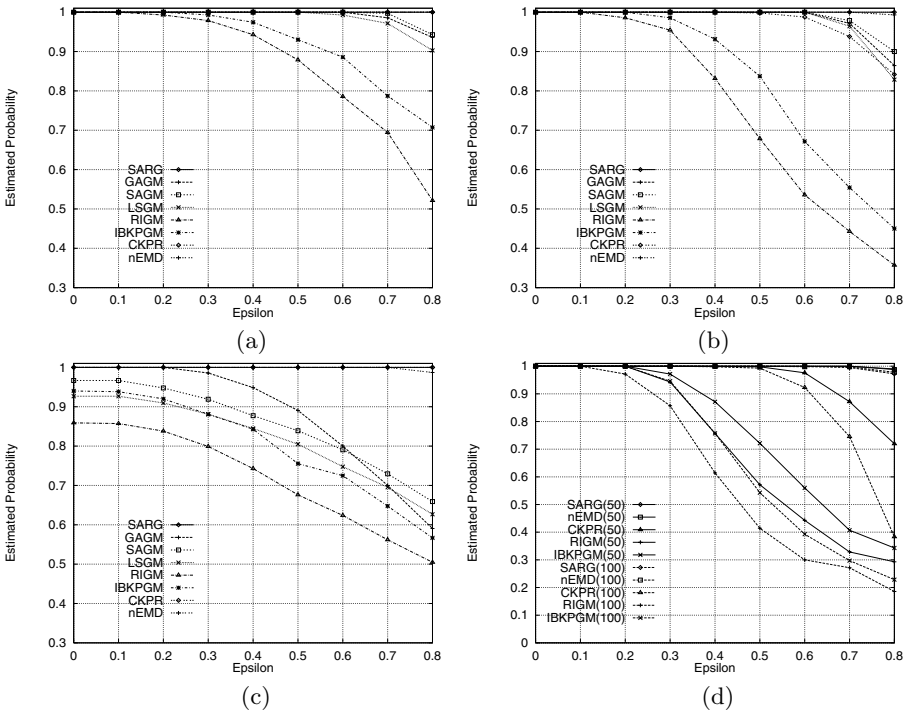


Fig. 3. Graph matching results under the independent noise condition with $(N^{\mathcal{G}_M}, N^{\mathcal{G}_I}, N_U, N_B) =$ (a) (10, 10, 3, 3), (b) (30, 30, 3, 3), (c) (15, 5, 5, 5), and (d) $(N^{\mathcal{G}_M}, N^{\mathcal{G}_I}, 3, 3)$, where $N^{\mathcal{G}_M} = N^{\mathcal{G}_I} = 50, 100$

and matching are all $O(N^{\mathcal{G}_M^2})$, while that of correcting process is $O(N^{\mathcal{G}_M} N^{\mathcal{G}_I})$, since it only requires the recomputation of $N^{\mathcal{G}_M} N^{\mathcal{G}_I}$ correspondence measures and the selection of the node having the maximum correspondence measure. And the number of candidate subgraphs denoted by $n(\mathcal{A}_T)$ is less than the product of the numbers of nodes in two graphs, that is, $n(\mathcal{A}_T) \leq \zeta \cdot N^{\mathcal{G}_M} N^{\mathcal{G}_I}$, where ζ is a constant that typically varies between 0.01 and 0.2. Thus the computational cost for the latter part is

$$T_2 = T_{\text{detection}} + T_{\text{correction}} + T_{\text{matchnig}} = O(\zeta \cdot N^{\mathcal{G}_M^3} N^{\mathcal{G}_I}) + O(\zeta \cdot N^{\mathcal{G}_M^2} N^{\mathcal{G}_I^2}) + O(\zeta \cdot N^{\mathcal{G}_M^3} N^{\mathcal{G}_I}) = O(N^{\mathcal{G}_M^3} N^{\mathcal{G}_I}).$$

In summary, the total computational cost required for the proposed SARG matching algorithm is

$$T_{\text{total}} = T_1 + T_2 = \max[O(N^{\mathcal{G}_M^2} N^{\mathcal{G}_I^2}), O(N^{\mathcal{G}_M^3} N^{\mathcal{G}_I})]. \tag{22}$$

5 Experimental Results

To evaluate the matching performance of the proposed algorithm, we have tested it on synthetic ARGs as in [16]. Synthetic ARGs were generated by the following

Table 1. The computational complexity of each graph matching algorithm in terms of the processing time for one pair graph matching

$(N^{\mathcal{G}_M}, N^{\mathcal{G}_I}, N_U, N_B)$	(15, 5, 5, 5)	(10, 10, 3, 3)	(30, 30, 3, 3)	(50, 50, 3, 3)	(100, 100, 3, 3)
nEMD	0.007s	0.009s	0.990s	10.24s	262.1s
CKPR	0.004s	0.005s	0.249s	4.34s	63.3s
SARG	0.001s	0.002s	0.145s	1.21s	14.9s

procedures: First, given fixed $(N^{\mathcal{G}_M}, N^{\mathcal{G}_I}, N_U, N_B)$, a reference graph \mathcal{G}_M was randomly generated, in which all attributes had a random number between 0 and 1. Then, an input graph was constructed using randomly generated permutation matrix \mathbf{P} . Next, independent noise matrices \mathbf{N}_i^U 's and \mathbf{N}_i^B 's were obtained by multiplying a uniformly distributed random variable on the interval $[-\frac{1}{2}, \frac{1}{2}]$ by the noise power $\epsilon \in [0, 1.0]$.

5.1 Independent Noise

In order to generate the independent noise, we fixed all noise correlation matrices such that $\forall \mathbf{C}_i^U = \mathbf{I}$ and $\forall \mathbf{C}_{lm}^B = \Delta_{lm}$. For the benchmarks, we have selected GAGM [8], SAGM [13], LSGM [14], RIGM [15] and IBKPGM [16] algorithms among various ARG matching algorithms in [16], since they showed better matching performance than others. Moreover, the performance was compared to CKPR [5] and nEMD [19]. The estimated probability of correct node-to-node matching was evaluated as a function of the noise magnitude ϵ . To reflect the graph matching performance in term of probability, for a given value of ϵ , we have done 300 trials for each graph matching algorithm.

Fig. 3 (a) and (b) summarize the matching results for the full graph matching when $(N^{\mathcal{G}_M}, N^{\mathcal{G}_I}, N_U, N_B) = (10, 10, 3, 3)$ and $(30, 30, 3, 3)$, respectively. It is noted that SARG outperforms other graph matching algorithms especially for large values of ϵ . And, generally, since nEMD, CKPR and SARG consider the NULL node explicitly, they showed superior subgraph matching performances than the others. However, as the noise power increased, the matching rate of nEMD decreased, while those of CKPR and SARG remained almost constant. In Fig. 3 (d), the performances of some algorithms are shown for $N^{\mathcal{G}_M} = N^{\mathcal{G}_I} = 50$ and 100. Due to the limitations on the computational cost and the memory, some algorithms were excluded for comparison. Actually, as the number of nodes increased, the graph matching performance became severely degraded. However, nEMD and SARG were very robust to the increase of the number of nodes, and SARG performed best.

5.2 Complexity Analysis

We have analyzed and compared the computational complexity of the proposed algorithm with those of nEMD and CKPR in terms of the processing time for one pair graph matching. We measured the processing time for one pair graph

matching by varying the number of nodes and the number of attributes of graphs, and calculated the average processing time after 500 trials per each condition. The results are presented in Table 5.2, where the processing time was evaluated in seconds. It is noted that the proposed algorithm is much faster than the others, especially for the graphs with a large number of nodes.

6 Conclusion

In order to match ARGs by subgraph isomorphism efficiently, in this paper, we proposed a novel stochastic attributed relational graph (SARG) matching technique using the stochastic analysis in the relation vector space, which embeds the global structure as well as the local structure centered at a specific node. The new concepts related to the correspondence, such as the stochastic neighborhood, the support flow, the correspondence measure and the similarity were defined in terms of the probability and the geometrical distribution of the attribute vectors in the relation vector space. The proposed SARG matching algorithm consists of 2 step procedures. In the first stage, a finite number of subgraphs were extracted from the test graph and ordered according to the correspondence measure to the model graph. Then, missing nodes for each candidate subgraphs were detected and the correspondences are reestablished by eliminating the effects of them. Finally, the refined subgraphs are matched to the model graph by measuring the similarity between them.

Experimental results on the synthetic ARGs demonstrated the robustness and efficiency of the proposed SARG matching algorithm. And it was also verified empirically that the proposed SARG matching algorithm was much faster than conventional graph-based algorithms.

Acknowledgement

This work has been supported in part by the ITRC (Information Technology Research Center) support program of Korean government and IIRC (Image Information Research Center) by Agency of Defense Development, Korea.

References

1. A. K. C. Wong and M. You, "Entropy and distance of random graphs with application to structural pattern recognition," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 7, no. 5, September 1985.
2. A. Sanfeliu, R. Alqu  zar, J. Andrade, J. Climent, F. Serratosa and J. Verg  s, "Graph-based representations and techniques for image processing and image analysis," *Pattern Recognition*, vol. 35, pp. 639-650, 2002.
3. F. Serratosa, R. Alqu  zar and A. Sanfeliu, "Function-described graphs for modelling objects represented by sets of attributes graphs," *Pattern Recognition*, vol. 36, pp. 781-798, 2003.

4. S. Z. Li, "Matching : invariant to translations, rotations and scale changes," *Pattern Recognition*, vol. 25, pp. 583-594, 1992.
5. W. J. Christmas, J. Kittler and M. Petrou, "Structural matching in computer vision using probabilistic relaxation," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 749-764, 1995.
6. B. T. Messmer and H. Bunke, "A new algorithm for error-tolerant subgraph isomorphism detection," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 20, no. 5, pp. 493-503, 1998.
7. B. T. Messmer and H. Bunke, "A decision tree approach to graph and subgraph isomorphism detection," *Pattern Recognition*, vol. 32, pp. 1979-1998, 1999.
8. S. Gold and A. Rangarajan, "A graduated assignment algorithm for graph matching," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 18, no. 4, pp. 377-388, April 1996.
9. W. H. Tsai and K. S. Fu, "Subgraph error-correcting isomorphisms for syntactic pattern recognition," *IEEE Trans. Systems Man and Cybernetics*, vol. 13, no. 1, pp. 48-62, January/February 1983.
10. Y. El-Sonbaty and M. A. Ismail, "A new algorithm for subgraph optimal isomorphism," *Pattern Recognition*, vol. 31, no. 2, pp. 205-218, 1998.
11. L. Herault, R. Horaud, F. Veillon and J. J. Niez, "Symbolic image matching by simulated annealing," *Proc. British Machine Vision Conference*, Oxford, pp. 319-324, 1990.
12. M. Krcmar and A. Dhawan, "Application of genetic algorithms in graph matching," *Proc. Int'l Conf. Neural Networks*, vol. 6, pp. 3872-3876, 1994.
13. B. J. van Wyk and M. A. van Wyk, "The spherical approximation graph matching algorithm," *Proc. Int'l Workshop on Multidisciplinary Design Optimization*, pp. 280-288, August 2000.
14. B. J. van Wyk and J. Clark, "An algorithm for approximate least-squares attributed graph matching," *Problems in Applied Mathematics and Computational Intelligence*, pp. 67-72, 2000.
15. M. A. van Wyk, T. S. Durrani and B. J. van Wyk, "A RKHS interpolator-based graph matching algorithm," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 988-995, 2002.
16. B. J. van Wyk and M. A. van Wyk, "Kronecker product graph matching," *Pattern Recognition*, vol. 36, pp. 2019-2030, 2003.
17. R. Nevatia and K. R. Babu, "Line extraction and description," *Computer Graphics and Image Processing*, vol.13, no.1, pp.250-269, July 1980.
18. B. G. Park, K. M. Lee, S. U. Lee, and J.H. Lee, "Recognition of partially occluded objects using probabilistic ARG-based matching," *Computer Vision and Image Understanding*, vol. 90, no. 3, pp. 217-241, June 2003.
19. D. H. Kim, I. D. Yun, and S. U. Lee, "A new attributed relational graph matching algorithm using the nested structure of earth mover's distance," *Proceedings of IEEE International Conference on Pattern Recognition*, Cambridge, UK, pp. 48-51, August 2004.

A New Similarity Measure for Random Signatures: Perceptually Modified Hausdorff Distance

Bo Gun Park, Kyoung Mu Lee, and Sang Uk Lee

School of Electrical Eng., ASRI, Seoul National University,
151-600, Seoul, Korea

gun@diehard.snu.ac.kr, kyoungmu@snu.ac.kr, sanguk@sting.snu.ac.kr

Abstract. In most content-based image retrieval systems, the low level visual features such as color, texture and region play an important role. Variety of dissimilarity measures were introduced for an uniform quantization of visual features, or a *histogram*. However, a cluster-based representation, or a *signature*, has proven to be more compact and theoretically sound for the accuracy and robustness than a histogram. Despite of these advantages, so far, only a few dissimilarity measures have been proposed. In this paper, we present a novel dissimilarity measure for a random signature, Perceptually Modified Hausdorff Distance (PMHD), based on Hausdorff distance. In order to demonstrate the performance of the PMHD, we retrieve relevant images for some queries on real image database by using only color information. The precision vs. recall results show that the proposed dissimilarity measure generally outperforms all other dissimilarity measures on an unmodified commercial image database.

1 Introduction

With an explosive growth of digital image collections, Content-Based Image Retrieval (CBIR) has been one of the most active and challenging problems in computer vision and multimedia applications [22][23]. There have been lots of image retrieval systems, which are based on the query-by-example scheme, including QBIC [20], PhotoBook [24], VisualSEEK [25], and MARS [26] etc. However, closing the gap between human perceptual concepts and low-level visual contents extracted by computer, is still one of ongoing problems. In order to deal with the semantic gap, many techniques have been introduced to improve visual features and similarity measures [22][4][27][28].

In most image retrieval system based on visual features, a *histogram* (or a *fixed-binning histogram*) is widely used as a visual feature descriptor due to its simple implementation and insensitivity to similarity transformation [4]. However, in some cases, the histogram based indexing methods fail to match perceptual dissimilarity [1]. The performance of retrieval system employing a histogram as a descriptor severely depends on the quantization process in feature space because a histogram is inflexible under various feature distribution representations. To overcome these drawbacks, a clustering based representation, *signature* (or

adaptive-binning histogram) has been proposed [1][3][15]. A signature compactly represents a set of clusters in feature space and the distribution of visual features. Therefore, it can reduce the complexity of representation and the cost of retrieval process. Once two sets of visual features based on a histogram or a signature, are given, it needs to determine how similar one is from the other. A number of different dissimilarity measures have been proposed in various areas of computer vision. Specifically for histograms, Jeffrey divergence, histogram intersection, χ^2 -statistics and so on have been known to be successful. However, these dissimilarity measures can not be directly applied to signatures. Rubner *et al.* [1] proposed a novel dissimilarity measure for matching signatures called the Earth Mover's Distance (EMD), which was able to overcome most of the drawbacks in histogram based dissimilarity measures and handled partial matches between two images. Dorado *et al.* [3] also used the EMD as a metric to compare fuzzy color signatures. However, the computational complexity of the EMD is very high compared to other dissimilarity measures. And Leow *et al.*[15] proposed a new dissimilarity measure, Weighted Correlation (WC) for signatures, which is more reliable than Euclidean distance and computationally more efficient than EMD. The performance of WC was generally better than EMD and comparable to other dissimilarity measures for image retrieval and image classification, but, in some cases, it was worse than the Jeffrey divergence (JD) [14].

In this paper, we propose a novel dissimilarity measure for comparison of random signatures, which is based on the Hausdorff distance. The Hausdorff distance is an effective metric for the dissimilarity measure between two sets of points [6][7][8][10], while insensitive to the characteristics changes of points. In this paper, we modify the general Hausdorff distance into the Perceptually Modified Hausdorff Distance (PMHD) in order to evaluate the dissimilarity between random signatures and to satisfy human perception. The experimental results on a real image database show that the proposed metric outperforms other dissimilarity measures.

2 A Visual Feature Descriptor: A Random Signature

In order to retrieve visually similar images to a query image using visual information, a proper visual feature descriptor should be extracted from an image. It has been proven that a signature can describe the feature distribution more efficiently than a histogram [1][3][15]. And a signature is appropriate for describing each image independently to other images in an image database.

In this paper, we represent an original image as a *random signature*, defined as

$$\mathcal{S} = \{(\mathbf{s}_i, w_i, \boldsymbol{\Sigma}_i) | i = 1, \dots, N\}, \quad (1)$$

where N is the number of clusters, \mathbf{s}_i is the mean feature vector of i -th cluster, w_i is the fraction of the features that belong to i -th cluster and $\boldsymbol{\Sigma}_i$ is the covariance matrix of i -th cluster. Variety of different clustering methods can be used to construct a random signature from a color image. In this paper, we used K-means clustering [12] to cluster visual features.

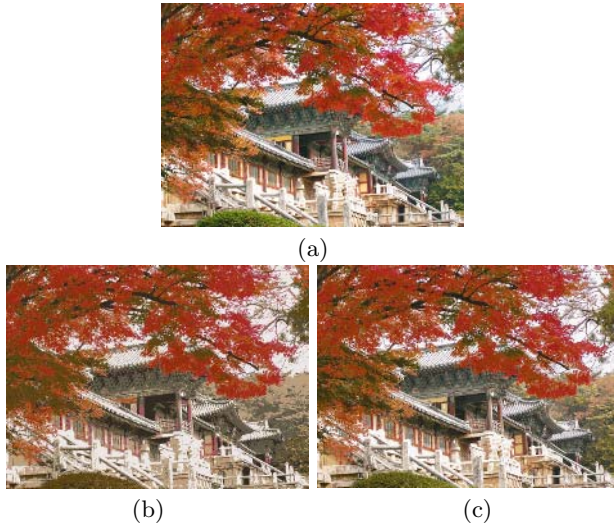


Fig. 1. Sample images quantized using K-mean clustering : (a) Original image with 256,758 colors, and quantized images based on a random signature with (b) 10 colors, and (c) 30 colors

Fig. 1 shows two sample images quantized by using a random signature with color information as a visual feature.

3 A Novel Dissimilarity Measure for a Random Signature

3.1 Hausdorff Distance

It has been shown that the Hausdorff distance (HD) is an effective metric for the dissimilarity measure between two sets of points in a number of computer vision literatures [6][7][8][9], while insensitive to the characteristics changes of points.

In this section, we briefly describe the Hausdorff distance(HD). More details can be found in [6][7][8][9]. Given two finite point sets, $\mathcal{P}_1 = \{p_1^1, \dots, p_N^1\}$ and $\mathcal{P}_2 = \{p_1^2, \dots, p_M^2\}$, the HD is defined as

$$\mathcal{D}_H = (\mathcal{P}_1, \mathcal{P}_2) = \text{Max}\{d_H(\mathcal{P}_1, \mathcal{P}_2), d_H(\mathcal{P}_2, \mathcal{P}_1)\}, \tag{2}$$

where

$$d_H(\mathcal{P}_1, \mathcal{P}_2) = \max_{p_1 \in \mathcal{P}_1} \min_{p_2 \in \mathcal{P}_2} \|p_1^1 - p_2^2\|, \tag{3}$$

and the function d_H is the directed HD between two point sets.

3.2 Perceptually Modified Hausdorff Distance

In this paper, we propose a novel dissimilarity, called Perceptually Modified Hausdorff Distance(PMHD) measure based on HD for comparison of random signatures.

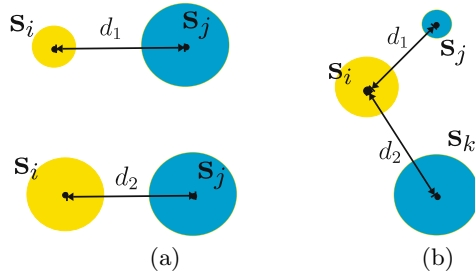


Fig. 2. An example of perceptual dissimilarity based on the densities of two color features

Given two random signatures $\mathcal{S}_1 = \{(s_i^1, w_i^1, \Sigma_i^1) | i = 1, \dots, N\}$, and $\mathcal{S}_2 = \{(s_j^2, w_j^2, \Sigma_j^2) | j = 1, \dots, M\}$, a novel dissimilarity measure between two random signatures is defined as

$$\mathcal{D}_{\mathcal{H}}(\mathcal{S}_1, \mathcal{S}_2) = \text{Max}\{d_H(\mathcal{S}_1, \mathcal{S}_2), d_H(\mathcal{S}_2, \mathcal{S}_1)\}, \tag{4}$$

where $d_H(\mathcal{S}_1, \mathcal{S}_2)$ and $d_H(\mathcal{S}_2, \mathcal{S}_1)$ are directed Hausdorff distances between two random signatures.

The directed Hausdorff distance is defined as

$$d_H(\mathcal{S}_1, \mathcal{S}_2) = \frac{\sum_i [w_i^1 \times \min_j \frac{d(s_i^1, s_j^2)}{\min(w_i^1, w_j^2)}]}{\sum_i w_i^1}, \tag{5}$$

where $d(s_i^1, s_j^2)$ is the distance between two visual features of the same type, s_i^1 and s_j^2 , which measures the difference between two features.

In (5), we divide the distance between two feature vectors by the minimum of two feature vectors' densities. Let's consider an example in Fig. 2(a). There are two pairs of feature vectors represented as circles centered at mean feature vectors. The size of each circle represents the density of the corresponding feature. If we compute only the geometric distance without considering the densities of two feature vectors, two distances d_1 and d_2 are equal. However, perceptually d_2 must be smaller than d_1 . Another example is given in Fig. 2(b). There are three feature vectors. d_1 is smaller than d_2 if we consider only the geometric distance regardless of the densities, however, it is perceptually justified that d_2 is smaller than d_1 . The desired distance should imply these observations. Therefore, we divide the geometric distance by the intersection of two feature vector's volume to match perceptual dissimilarity.

In the result, PMHD is insensitive to the characteristics changes of mean features in a signature and theoretically sound for involving human intuition and perception in the metric.

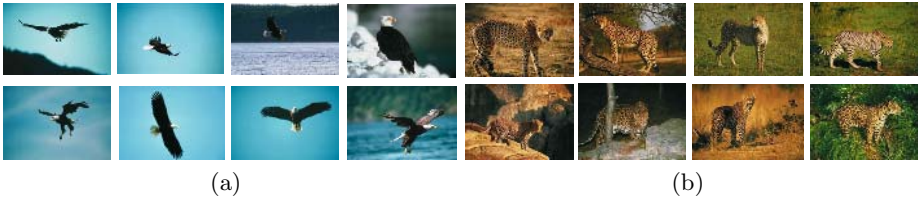


Fig. 3. Example query images from four categories in the Corel database. (a) Eagle, (b) Cheetah.

3.3 Partial PMHD Metric for Partial Matching

If a user is interested in only a part of images or requires to retrieve partially similar images, a global descriptor is not appropriate for such task. Like a histogram, a signature is also a global descriptor of a whole image. And the proposed distance for random signatures in (4) can include possible outliers by employing the summation operator over all distances. As indicated in [1][5][9], this kind of distance can not cope with occlusion and clutter in image retrieval or object recognition. In order to handle partial matching, Huttenlocher *et al.* [6] proposed partial HD based on ranking, which measures the difference between portion of sets of points. And Azencott *et al.* [8] further modified rank based partial HD by order statistics. But, these distances were shown to be sensitive to the parameter changes. In order to address these problems, Sim *et al.* [9] proposed two robust HD measures, M-HD and LTS-HD, based on the robust statistics such as M-estimation and Least Trimmed Square(LTS). Unfortunately, they are not appropriate for image retrieval system because they are computationally too complex to search a large database.

In this section, we explicitly remove outliers in the proposed distance to address partial matching problem. Let us define outlier test function as

$$f(i) = \begin{cases} 1, & \min_j \frac{d(s_i^1, s_j^2)}{\min(w_i^1, w_j^2)} < D_{th}, \\ 0, & \text{otherwise,} \end{cases} \tag{6a}$$

$$\tag{6b}$$

where D_{th} is a pre-specific threshold for the outlier detection.

Then we compute two directed Hausdorff distance, $d_H^a(\mathcal{S}_1, \mathcal{S}_2)$ and $d_H^p(\mathcal{S}_1, \mathcal{S}_2)$, defined as

$$d_H^a(\mathcal{S}_1, \mathcal{S}_2) = \frac{\sum_i w_i^1 \times \min_j \frac{d(s_i^1, s_j^2)}{\min(w_i^1, w_j^2)}}{\sum_i w_i^1},$$

$$d_H^p(\mathcal{S}_1, \mathcal{S}_2) = \frac{\sum_i w_i^1 \times \min_j \frac{d(s_i^1, s_j^2)}{\min(w_i^1, w_j^2)} \times f(i)}{\sum_i w_i^1 \times f(i)}. \tag{7}$$

Now, let us modify the directed Hausdorff distance in (5) as

$$d_H(\mathcal{S}_1, \mathcal{S}_2) = \begin{cases} d_H^a(\mathcal{S}_1, \mathcal{S}_2), & \frac{\sum_i w_i^1 \times f(i)}{\sum_i w_i^1} < P_{th}, \\ d_H^p(\mathcal{S}_1, \mathcal{S}_2), & \text{otherwise,} \end{cases} \quad (8a)$$

$$(8b)$$

where P_{th} is a pre-specific threshold for the control of a faction of information loss.

4 Experimental Results

4.1 The Database and Queries

To evaluate the performance of the proposed metric, several experiments have been conducted on a real database with a color feature as a visual feature. We used 5,200 images selected from commercially available Corel color image database without any modification. There are 52 semantic categories, each of them containing 100 images. Among those, we have chosen four sets of query data, Cheetah, Eagle, Pyramids and Royal guards. Some example images in the queries are given in Fig. 3. In this experiment, we used all images in these four categories as a query. As we note in Fig. 3, grouping of images to different categories were not based on the color information. Nonetheless, in wide sense, it was considered that all images in the same category were considered as the relevant images or correct answers based on the color information. We computed a precision and recall pair to all query categories, which is commonly used as the retrieval performance measurement [11]. Precision P and recall R are defined as

$$P = r/n, \quad R = r/m, \quad (9)$$

where r is the number of retrieved relevant images, n is the total number of retrieved images, and m is the total number of relevant images in the whole database. Precision P measures the accuracy of the retrieval and recall R measures the robustness of the retrieval performance.

In this paper, we used only color feature as a visual feature. Thus we consider three different distances for $d(\mathbf{s}_i^1, \mathbf{s}_j^2)$ in (5) : the Euclidean distance, the CIE94 color difference, and the Mahalanobis distance. In order to guarantee that the distance is perceptually uniform, the CIE94 color difference equation is used instead of the Euclidean distance in CIELab color space [17]. And the Mahalanobis distance explicitly considers the distribution of color features after clustering process [16]. Three distances are defined as follows.

(i) Euclidean distance :

$$d_E(\mathbf{s}_i^1, \mathbf{s}_j^2) = \sum_{k=1}^3 [\mathbf{s}_i^1(k) - \mathbf{s}_j^2(k)]^{1/2}, \quad (10)$$

where $\mathbf{s}_i(k)$ is the k -th element in the feature vector \mathbf{s}_i .

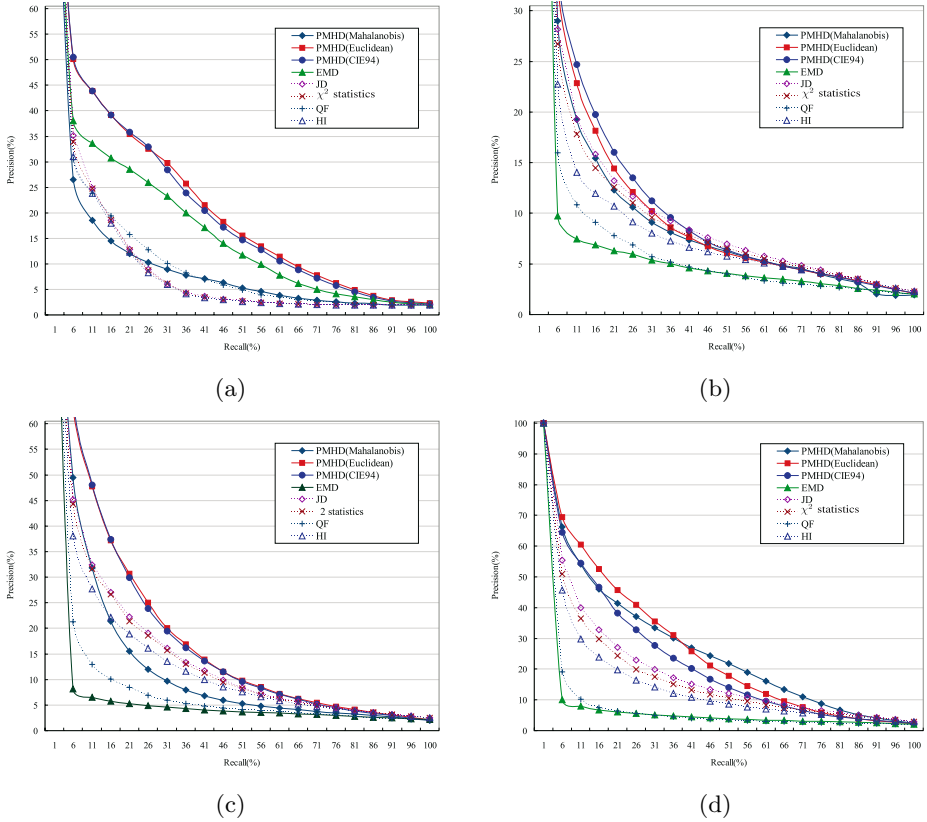


Fig. 4. Precision-recall curves for various dissimilarity measures on four query categories : (a) Eagle, (b) Cheetah, (c) Pyramids, and (d) Royal guards

(ii) CIE94 color difference equation :

$$d_{CIE94}(s_i^1, s_j^2) = [(\frac{\Delta L^*}{k_L S_L})^2 + (\frac{\Delta C^*}{k_C S_C})^2 + (\frac{\Delta H^*}{k_H S_H})^2]^{1/2}, \tag{11}$$

where ΔL^* , ΔC^* and ΔH^* are the differences in lightness, chroma, and hue between s_i^1 and s_j^2 .

(iii) Mahalanobis distance :

$$d_M(s_i^1, s_j^2) = (s_i^1 - s_j^2)^T \Sigma_i^{-1} (s_i^1 - s_j^2). \tag{12}$$

4.2 Retrieval Results for Queries

The performance of the proposed PMHD was compared with five well-know dissimilarity measures, including *Histogram Intersection(HI)*, χ^2 -*statistics*, *Jeffrey-*

Divergence(JD) and *Quadratic Form(QF) distance* for the fixed binning histogram, and *EMD* for the signature. Let H_1 and H_2 represent two color histograms or signatures. Then, these dissimilarity measures are defined as follows.

– *Histogram Intersection(HI)* [18] :

$$d(H_1, H_2) = 1 - \sum_i \min(h_i^1, h_i^2) / \sum_i h_i^2, \tag{13}$$

where h_i^j is the number of elements in i -th bin of H_j .

– χ^2 -statistics :

$$d(H_1, H_2) = \sum_i (h_i^1 - m_i)^2 / m_i, \tag{14}$$

where $m_i = (h_i^1 + h_i^2) / 2$.

– *Jeffrey-Divergence(JD)* [14] :

$$d(H_1, H_2) = \sum_i (h_i^1 \log \frac{h_i^1}{m_i} + h_i^2 \log \frac{h_i^2}{m_i}), \tag{15}$$

where again $m_i = (h_i^1 + h_i^2) / 2$.

– *Quadratic Form(QF) distance* [19][20] :

$$d(H_1, H_2) = \sqrt{(H_1 - H_2)^T A (H_1 - H_2)}, \tag{16}$$

where A is a similarity matrix. A encodes the cross-bin relationships based on the perceptual similarity of the representative colors of the bins.

– *EMD* [1] [13] :

$$d(H_1, H_2) = \sum_{i,j} g_{ij} d_{ij} / \sum_{i,j} g_{ij}, \tag{17}$$

where d_{ij} denotes the dissimilarity between i -th bin and j -th bin, and g_{ij} is the optimal flow between two distributions. The total cost $\sum_{i,j} g_{ij} d_{ij}$ is

minimized subject to the constraints,

$$\begin{aligned} g_{ij} &\geq 0, \sum_i g_{ij} \leq h_j^2, \sum_j g_{ij} \leq h_i^1, \\ \sum_{i,j} g_{ij} &= \min(\sum_i h_i^1, \sum_j h_j^2). \end{aligned} \tag{18}$$

As reported in [13], EMD yielded very good retrieval performance for the small sample size, while JD and χ^2 performed very well for the larger sample sizes. Leow *et al.* [15] proposed the novel dissimilarity measure, Weighted Correlation(WC) which can be used to compare two histograms with different

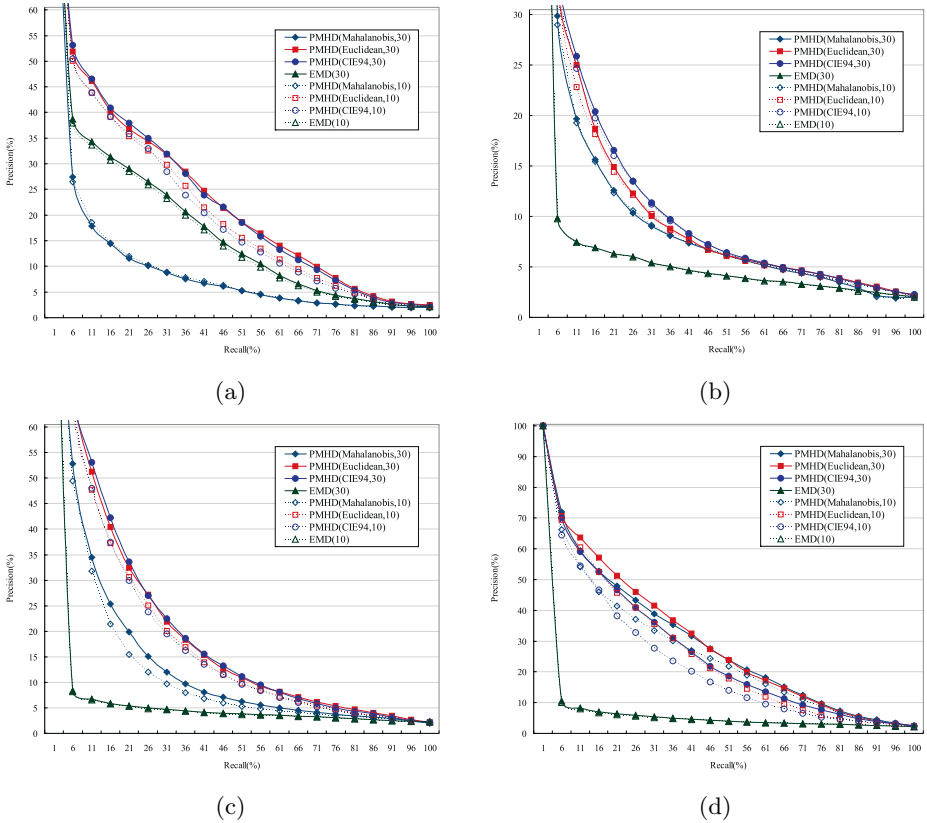


Fig. 5. Comparison of the retrieval performance for varying the number of color features in a signature : (a) Eagle, (b) Cheetah, (c) Pyramids, and (d) Royal guards

binnings. In the image retrieval, the performance of WC was comparable to other dissimilarity measures, however, JD always outperform WC. And, in this paper, we evaluated only the performance of JD. In order to represent a color image as a fixed histogram representation, the RGB color space was uniformly partitioned into $10 \times 10 \times 10 = 1000$ color bins. And a color was quantized to the mean centroid of the cubic bin. While, as mentioned in Section 2, a random signature was extracted by applying K-means clustering. To compare the performance of the signature based dissimilarity with other fixed histogram based ones, the quantization level was matched by clustering a color image into only 10 color feature clusters. The mean color quantization error of the $10 \times 10 \times 10$ -bin histogram is 5.99 CIE94 units and that of quantized image based on a random signature containing 10 color feature vectors was 5.26 CIE94 units. It is noted that the difference between two quantized image errors are smaller than the perceptibility threshold of 2.2 CIE94 units [21], where two colors are perceptually indistinguishable [15].

The retrieval performance results of the proposed metric and other dissimilarity methods are summarized by the precision-recall in Fig. 4. It is noted that the proposed PMHD dissimilarity measure significantly outperformed other dissimilarity measures for all query images. The performance of PMHD is, on average, 20–30% higher than the second highest precision rate over the meaningful recall values. And the performance of PMHD with Euclidean distance is almost the same as that of PMHD with CIE94, and usually performed best in the image retrieval. It is somewhat surprisingly noted that EMD performed poorer than other dissimilarity measures in all query categories except “Eagle” query category. This performance is not coincident with the results reported in [13] and [1], where EMD performed very well for the small sample sizes and the compact representation but not so well for large sample sized and wide representation. As indicated in [15], the image size, the number of color features in a signature and the ground distance may degrade the whole performance of EMD. However, as mentioned before, we only used a signature with 10 color features in this experiment, which is a very compact representation. We note that the large image size of 98,304 pixels or so and the Euclidean ground distance may severely degrade the performance of EMD.

4.3 Dependency on the Number of Color Features in a Signatures

In general, the quantization level of a feature space, that is, the number of clusters in a signature or the number of bins in the fixed histogram, has an important effect on the overall image retrieval performance. In order to investigate the retrieval performance dependency on the quantization level, we compared the retrieval performance of the proposed method according to the number of color features in a signature, which varied for 10 and 30. The mean color error of the quantized image based on a random signature with 30 color feature vectors is 3.38 CIE94 units, which is significantly smaller than 5.26 CIE94 units in the case of a random signature with 10 color feature vectors. Fig. 1 shows two sample images quantized using K-means clustering, which were quantized by 10 colors and 30 colors each. It is noted that the quantized image based on a random signature with 30 color features is almost indistinguishable from the original image, which contains 256,758 color features.

Fig. 5 plots the precision-recall curves of the image retrieval results for varying the number of color features in a signature. We compared the retrieval performance of the proposed PMHD with EMD, since it is the only dissimilarity measure applicable to signatures. The precision rate of EMD does not vary significantly as the number of color features of a signature increased, as depicted in Fig. 5. However, the precision rate of PHMD with 30 color features is slightly higher than that of PMHD with 10 color features. From this result, it can be expected that the performance of the proposed PMHD becomes higher as the quantization error decreases. Moreover, this implies that PMHD performs best for the large sample sizes as well as the compact representation.

5 Conclusion

In this paper, we proposed a novel dissimilarity measure for random signatures, Perceptually Modified Hausdorff Distance (PMHD) based on Hausdorff distance. PMHD is insensitive to the characteristics changes of mean features in a signature and theoretically sound for human intuition and perception of the metric. The extensive experimental results on a real database showed that the proposed PMHD outperformed other dissimilarities. The retrieval performance of the PMHD is, on average, 20–30% higher than the second highest precision rate. In this paper, we used only color information, which was shown to be inappropriate to close the semantic gap without using texture information, multi-resolution representation, relevance feedback, and so on. Thus, combining texture information and representing signature in multi-resolution framework will be our future work.

Acknowledgement

This work has been supported in part by the ITRC (Information Technology Research Center) support program of Korean government and IIRC (Image Information Research Center) by Agency of Defense Development, Korea.

References

1. Y. Rubner and C. Tomasi, *Perceptual metrics for image database navigation*, Kluwer Academic Publisher, January 2001.
2. G. Qiu and K.M. Lam, "Frequency layered color indexing for content-based image retrieval," *IEEE Trans. Image Processing*, vol.12, no.1, pp.102-113, January 2003.
3. A. Dorado and E. Izquierdo, "Fuzzy color signature," *IEEE Int'l Conference on Image Processing*, vol.1, pp.433-436, 2002.
4. A.W.M. Smeulders *et al.*, "Content-based image retrieval at the end of the early years," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.22, no.12, pp.1349-1380, December 2000.
5. V. Gouet and N. Boujemaa, "About optimal use of color points of interest for content-based image retrieval," *Research Report RR-4439*, INRIA Rocquencourt, France, April 2002.
6. D.P. Huttenlocher, G.A. Klanderman, and W.J. Rucklidge "Comparing images using the Hausdorff distance," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol.15, no.9, pp.850-863, September 1993.
7. M.P. Dubuisson and A.K. Jain, "A modified Hausdorff distance for object matching," *Proceedings of IEEE International Conference on Pattern Recognition*, pp.566-568, October 1994.
8. R. Azencott, F. Durbin, and J. Paumard, "Multiscale identification of building in compressed large aerial scenes," *Proceedings of IEEE International Conference on Pattern Recognition*, vol.2, pp.974-978, Vienna, Austria, 1996.
9. D.G. Sim, O.K. Kwon, and R.H. Park, "Object matching algorithms using robust Hausdorff distance measures," *IEEE Trans. Image Processing*, vol.8, no.3, pp.425-428, March 1999.

10. S.H.Kim and R.H.Park, "A novel approach to video sequence matching using color and edge features with the modified Hausdorff distance," in *Proc. 2004 IEEE Int. Symp. Circuit and Systems*, Vancouver, Canada, May 2004.
11. Alberto Del Bimbo, *Visual information retrieval*, Morgan Kaufmann Publishers Inc., San Francisco, CA, 1999.
12. R.O.Duda, P.E.Har, and D.G.Stork, *Pattern classificatoin*, Wiley & Sons Inc., New York, 2001.
13. J.Puzicha, J.M.Buhmann, Y.Rubner, and C.Tomasi, "Empirical evaluation of dissimilarity measures for color and texture," *Proceedings of IEEE International Conference on Computer Vision*, pp.1165-1173, 1999.
14. J.Puzicha, T.Hofmann, and J.Buhmann, "Nonparametric similarity measures for unsupervised texture segmentation and image retrieval," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp.267-272, June 1997.
15. W.K.Leow and R.Li, "The analysis and applications of adaptive-binning color histograms," *Computer Vision and Image Understanding*, vol.94, pp.67-91, 2004.
16. F.H.Imai, N.Tsumura, Y.Miyake, "Perceptual color difference metric for complex images based on Mahalanobis distance," *Journal of Electronic Imaging*, vol.10, no.2, pp.385-393, 2001.
17. K.N.Plataniotis and A.N.Venetsanopoulos *Color image processing and applications*, Springer, New York 2000.
18. M.Swain and D.Ballard, "Color indexing," *International Journal of Computer Vision*, vol.7, no.1, pp.11-32, 1991.
19. J.Hafner, H.S.Sawhney, W.Equitiz, M.Flickner, and W.Niblack, "Efficient color histogram indexing for quadratic form distance functions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, , vol.17, no.7, pp.729-735, July 1995.
20. M.Flickner, H.Sawhney, W.Niblack, J.Ashley, Q.Huang, B.Dom, M.Gorkani, J.Hafner, D.Lee, D.Petkovic, D.Steele, P.Yanker, "Query by image and video content : QBIC system," *IEEE Comput.*, vol.29, no.9, pp.23-32, 1995.
21. T.Song and R.Luo, "Testing color-difference formulae on complex images using a CRT monitor," in *Proc. 8th Color Imaging Conference*, 2000.
22. Y.Rui, T.S.Huang, and S.F.Chang, "Image retrieval : Current techniques, promising directions and open issues," *Journal of Visual Communication and Image Representation*, 1999.
23. W.Y.Ma and H.J.Zhang, "Content-based image indexing and retrieval," *Handbook of Multimedia Computing*, CRC Press, 1999.
24. A.Pentland, R.W.Picard, and S.Sclaroff, "Photobook : Content-based manipulation of image databases," *International Journal of Computer Vision*, vol.18, no.3, pp.233-254, 1996.
25. J.R.Smith and S.F.Chang, "VisualSEEK : A fully automated content-based image query system," *ACM Multimedia*, Boston MA, 1996.
26. Y.Rui, T.Huang, and S.Mehrotra, "Content-based image retrieval with relevance feedback in MARS," *IEEE Int'l Conference on Image Processing*, 1997.
27. T.Wang, Y.Rui, and J.G.Sun, "Constraint based region matching for image retrieval," *International Journal of Computer Vision*, vol.56, no.1/2, pp.37-45, 2004.
28. K.Tieu and P.Viola, "Boosting image retrieval," *International Journal of Computer Vision*, vol.56, no.1/2, pp.17-36, 2004.

Tracking of Linear Appearance Models Using Second Order Minimization

Jose Gonzalez-Mora, Nicolas Guil, and Emilio L. Zapata

Computer Architecture Department, University of Malaga

Abstract. The visual tracking of image regions is a research area of great interest within the computer vision community. One issue which has received quite attention in the last years has been the analysis of tracking algorithms which could be able to cope with changes in the appearance of the target region. Probably one of the most studied techniques proposed to model this appearance variability is that based on linear subspace models. Recently, efficient algorithms for fitting these models have been developed too, in many cases as an evolution of well studied approaches for the tracking of fixed appearance images.

Additionally, new methods based on second order optimizers have been proposed for the tracking of targets with no appearance changes. In this paper we study the application of such techniques in the design of tracking algorithms for linear appearance models and compare their performance with three previous approaches. The achieved results show the efficiency of the use of second-order minimization in terms of both number of iterations required for convergence and convergence frequency.

1 Introduction

Visual object tracking is a key element in many important applications of computer vision such as face analysis, mosaic reconstruction, augmented reality or advanced interfaces for human computer interaction. For many applications the computational simplicity of working with 2D patches makes it an interesting option before other alternatives as recovering the motion of 3D models.

In the majority of real world applications, the target region will suffer from changes in its appearance due to several factors: changes in the illumination, occlusions or the possibility of within-class object aspect variations itself (e.g. facial expressions). When this variability is not relevant for the applications (e.g., face pose estimation), we can employ algorithms which use invariant descriptions or robust metrics to make the tracking performance independent from aspect changes. However, if the appearance information needs to be used (e.g., face expression analysis) then algorithms capable of modeling aspect changes must be used. These algorithms must contain enough information to represent the underlying degrees of freedom of the appearance of the imaged object.

Linear models are a well known method for representing appearance due its computational efficiency and simplicity. They consist in a basis of template images whose linear combination can be used to approximate the object viewed under different conditions (e.g. illumination, expressions). In general, the

relationship between the image and the subspace which contains all the possible instances of the target region is not linear; when constructing the image basis, we will look for a linear approximation of this subspace, using tools as Principal Component Analysis (PCA). PCA provides a efficient way to reduce the dimensionality of the input training data.

The process of tracking consists of moving, deforming and adapting the appearance of the target representation to fit it in a image, minimizing the pixels difference. Traditional tracking approaches include techniques such as normalized correlation or template matching. In particular, we will focus here on template matching based algorithms using sum-of-squares (SSD) differences because they have received a big amount of attention in the last years and efficient approaches have also been developed [6,7,9].

Recently, Benhimane and Malis [2] developed an optimization scheme for template matching which works with a second order approximation.

In this paper, we have reformulated the second order optimization within an iterative tracking process where the role of image and template are clearly established and the incremental characteristic of the tracking is explicitly shown. In addition, we have applied the second order optimization approach to the tracking of linear models of appearance. This is achieved by introducing the second order formulation in three first order tracking algorithms: independent optimization of pose and appearance, simultaneous pose and appearance optimization, and projected out optimization.

The paper is organized as follows. In the next section we introduce a new formulation for the tracking problem with fixed appearance using second order minimization. In Section 3, several well known approaches for tracking of images, using linear optimization and variable appearance are shown. In Section 4 second order minimization is applied to the techniques presented in previous section. In Section 5 we describe some experiments that show the efficiency of the second order optimization to track linear appearance models. Finally, in Section 6 some conclusions and directions of future work are given.

2 Tracking Fixed Appearance Images

Let $I(x, t)$ denote the brightness value at the location $x = (x, y)$ in an image acquired at time t and $R = (x_1, x_2, \dots, x_N)$ the set of N image locations which define a target region. The relative motion between object and camera in the tracking process causes a transformation in the target region. We can represent this deformation by a motion model $W(x, p)$ parametrized by $p = (p_1, p_2, \dots, p_n)^T$, with W differentiable both in x and p .

If we assume that variations in brightness values are only caused by the target motion (the image constancy assumption holds for all pixel in R), the tracking process can be formulated as minimizing the following least squares function with respect to p :

$$\sum_x \|T(x) - I(W(x, p))\|^2 \quad (1)$$

In general, optimizing this expression is a difficult task, which cannot be linearly solved; however, in the case of tracking applications we have the possibility of using the continuity of motion to estimate a starting point for the minimization and develop smart iterative approaches to find the optimal parameters [5]. Furthermore, low computational cost schemes have been developed, reformulating the image alignment process in such a way that a great part of the needed computations can be done offline [4].

2.1 Inverse Compositional SSD Image Tracking

Inverse compositional algorithm [3] is one of the most used approaches to image tracking due to its simplicity, generality and efficiency. It iteratively minimizes an expression similar to equation (1):

$$\sum_x \|T(W(x, \Delta p)) - I(W(x, p))\|^2 \quad (2)$$

and then updates the warp parameters as:

$$W(x, p) \leftarrow W(x, p) \circ W(x, \Delta p)^{-1} \quad (3)$$

Making a Taylor series expansion of order 1 of (2) about $p = 0$ we have:

$$\sum_x \|T(W(x, 0)) + \nabla T \frac{\partial W}{\partial p} \Delta p - I(W(x, p))\|^2 \quad (4)$$

Assuming that $W(x, 0)$ is the identity warp, the solution to this least-squares problem is:

$$\Delta p = -H^{-1} \sum_x J(x)e(x), \quad (5)$$

with:

- $e(x) = T(W(x, 0)) - I(W(x, p))$ the captured image error.
- $J = \nabla T \frac{\partial W}{\partial p}$ the so called "steepest-descent" images, which represent "motion modes" as they relate the changes in brightness induced by the motion represented by the corresponding motion parameters. ∇T is the image template spatial gradient and $\frac{\partial W}{\partial p}$ the warp function Jacobian.
- $H = \sum_x J^T(x)J(x)$ the newton approximation of the Hessian matrix.

As we can see, the J and H terms can be precomputed and only the shown matrix multiplication must be done in each iteration.

2.2 Second Order SSD Image Tracking

Recently, Benhimane and Malis, [2], proposed a efficient second-order minimization method applicable to image tracking, reporting some improvements over previous approaches. In this section a new formulation of this technique more suitable to be applied to an inverse compositional algorithm is shown.

The problem formulation is the same we have seen for the inverse compositional algorithm. That is, expression 2 must be minimized in order to get the deformation parameters which align the reference template with the input image. However, to achieve a higher convergence order, a second order approximation of the deformed template is developed:

$$T(W(x, \Delta p)) \approx T(W(x, 0)) + \frac{\partial T(W(x, p))}{\partial p} \Big|_{p=0} \Delta p + \frac{1}{2} \Delta p^T \frac{\partial^2 T(W(x, p))}{\partial p^2} \Big|_{p=0} \Delta p \tag{6}$$

If this expression were substituted in (2), then the second order partial derivative should be calculated during the minimization process. However, this computation can be avoided using the Taylor series of the Jacobian of $T(W(x, p))$ at $p = 0$ and evaluated at Δp :

$$\begin{aligned} \frac{\partial T(W(x, \Delta p))}{\partial p} &\approx \\ &\approx \frac{\partial T(W(x, p))}{\partial p} \Big|_{p=0} + \frac{\partial^2 T(W(x, p))}{\partial p^2} \Big|_{p=0} \Delta p \end{aligned} \tag{7}$$

Working out $\frac{\partial^2 T(W(x, p))}{\partial p^2} \Delta p$ and substituting in (6):

$$T(W(x, \Delta p)) \approx T(W(x, 0)) + \frac{1}{2} \left(\frac{\partial T(W(x, p))}{\partial p} \Big|_{p=0} + \frac{\partial T(W(x, \Delta p))}{\partial p} \right) \Delta p \tag{8}$$

In this last expression, a term, $\frac{\partial T(W(x, \Delta p))}{\partial p}$, needs to be evaluated at the unknown point Δp (that is precisely the value to be obtained as result of the iteration). This calculation can be approximated by decomposing the partial derivative in the following way:

$$\begin{aligned} \frac{\partial T(W(x, \Delta p))}{\partial p} &= \frac{\partial T(W(W(x, \Delta p^{-1} \circ p), \Delta p))}{\partial p} = \\ &= \nabla_x T(W(x, \Delta p)) \frac{\partial W(x, \Delta p^{-1} \circ p)}{\partial p} \Big|_{p=\Delta p} = \\ &= \nabla_x T(W(x, \Delta p)) \frac{\partial W(x, p)}{\partial p} \Big|_{p=0} \frac{\partial (\Delta p^{-1} \circ p)}{\partial p} \Big|_{p=\Delta p} \end{aligned} \tag{9}$$

where

- $\nabla_x T(x, \Delta p)$ is the spatial gradient of the template image. In the inverse compositional formulation Δp is the deformation that aligns the template image with the current input image (2). Then

$$T(W(x, \Delta p)) \approx I(W(x, p_c))$$

and the following approximation can be carried out:

$$\nabla T(W(x, \Delta p)) \approx \nabla I(W(x, p_c))$$

- $\frac{\partial W(x,p)}{\partial p} \Big|_{p=0}$ is the warping Jacobian function evaluated at the identity.
- The term $\frac{\partial(\Delta p^{-1} \circ p)}{\partial p} \Big|_{p=\Delta p}$ can be approximated by the identity matrix $I_{n \times n}$ for several parametrization. Thus, in *Lie Algebra Parametrization* a first order approximation of the exponential parameters composition is $p^{-1} \circ p \approx p^{-1} + p$ [2], from where $\frac{\partial p^{-1} \circ p}{\partial p} \approx I_{n \times n}$. We have checked that this approach is also true for the linear parametrization proposed in [3].

Therefore, when using a valid parametrization, we can choose a better linear approximation for $T(W(x, \Delta p))$, instead of the Taylor expansion of first order, given by:

$$T(W(x, \Delta p)) \approx T(W(x, 0)) + \frac{1}{2}(\nabla_x T + \nabla_x I(W(x, p))) \frac{\partial W(x, p)}{\partial p} \Big|_{p=0} \Delta p \quad (10)$$

Finally, substituting this approximation in (2):

$$E(\Delta p) \approx \sum \|T(x) - I(W(x, p_c)) + \frac{1}{2}(\nabla_x T + \nabla_x I(W(x, p))) \frac{\partial W(x, p)}{\partial p} \Big|_{p=0} \Delta p\|^2 \quad (11)$$

The resulting optimal parameter can be obtaining by applying:

$$\Delta p = H_{esm}^{-1} \sum_x J_{esm}^T * e_{esm}(x) \quad (12)$$

where:

- $e_{esm}(x) = T(x) - I(W(x, p_c))$
- $J_{esm} = \frac{1}{2}(\nabla_x T(x) + \nabla_x I(W(x, p_c))) \frac{\partial W(x, p)}{\partial p} \Big|_{p=0}$
- $H_{esm} = J_{esm}^T * J_{esm}$

3 Linear Image Models Tracking Formulation

In this section, several previous algorithms for the tracking process of images with variable appearance using linear models are presented. Now, the target appearance is represented using a reference template $T(x)$ and a linear basis of images $\{A_i\}, i = 1, \dots, m$ modeling the changes which can occur in the object appearance. The tracking process can be expressed as:

$$\sum_x \|T(x) + \sum_i \lambda_i A_i - I(W(x, p))\|^2 \quad (13)$$

where p and $\lambda = (\lambda_1, \lambda_2, \dots, \lambda_m)^T$ are the pose and appearance parameter vectors respectively.

Hager and Belhumeur implemented this approach for the robust tracking of targets under variable illumination [6]; Black and Jepson used this algorithm for general appearance variation [7].

In the following we will compare three different approaches to optimize this expression.

3.1 Independent Optimization of Pose and Appearance

This solution tries to optimize separately warp and illumination parameters. Based on this idea several algorithms can be found in the bibliography, as the proposed in [8], that implements the inverse additive paradigm. However, in this paper an inverse compositional approach has been used. Then, the error is calculated as follows:

$$E(\Delta\lambda, \Delta p) = \sum_x \|T(W(x, \Delta p)) + \sum_m \Delta\lambda_i A_i(W(x, \Delta p)) - I(W(x, p)) - \sum_i \lambda_i A_i(W(x, p))\|^2 \tag{14}$$

The method assumes λ constant and compute the minimum of $E(\Delta\lambda, \Delta p)$ w.r.t Δp , using the approximate expression obtained from the Taylor expansion on $T(W(x, p))$:

$$E(\Delta p) = \sum_x \|e(x) - \nabla T \frac{\partial W}{\partial p} \Delta p\|^2 \tag{15}$$

resulting the following pose parameter increment:

$$\Delta p = H_{iic}^{-1} \sum_x J_{iic}^T e_{iic}(x),$$

where:

- $e_{iic}(x) = T(x) - I(W(x, p)) - \sum_i \lambda_i A_i(W(x, p))$
- $J_{iic} = \nabla T \frac{\partial W}{\partial p}$
- $H_{iic} = \sum_x J_{iic}^T(x) J_{iic}(x)$

This expression of pose computing is similar to that with no appearance changes, except for the compensation of appearance changes introduced in the error value $e_{iic}(x)$.

Now, minimizing over $\Delta\lambda$ and assuming p constant:

$$E(\Delta\lambda) = \sum_x \|e(x) + \sum_i \Delta\lambda_i A_i(x)\|^2 \tag{16}$$

from where we get:

$$\Delta\lambda = (A^T A)^{-1} \sum_x A^T e(x)$$

using the same expression error as above. If $\{A_i\}, i = 1, \dots, m$ is an orthonormal basis it can be simplified as $A^T A = I$.

Finally, we compose the parameters as: $p = p \circ \Delta p^{-1}, \lambda = \lambda - \Delta \lambda$

3.2 Simultaneous Pose and Appearance Optimization

Accurate optimization requires to find the point where pose and appearance parameters make zero the derivative of the registration error:

$$E(\Delta \lambda, \Delta p) = \sum_x \|T(W(x, \Delta p)) + \sum_i (\lambda_i + \Delta \lambda_i) A_i(W(x, \Delta p)) - I(W(x, p))\|^2 \tag{17}$$

Now, optimization is carried out simultaneously with respect to both Δp and $\Delta \lambda$ at each iteration. Performing a first order Taylor expansion on $T(W(x, \Delta p))$ and $A_i(W(x, \Delta p))$ in (17) and assuming that $W(x, 0)$ is the identity warp, we have:

$$E(\Delta \lambda, \Delta p) = \sum_x \|T(x) + \nabla T \frac{\partial W}{\partial p} \Delta p + \sum_i (\lambda_i + \Delta \lambda_i) (A_i(x) + \nabla A_i \frac{\partial W}{\partial p} \Delta p) - I(W(x, p))\|^2 \tag{18}$$

Neglecting second order terms, the above expression simplifies to:

$$E(\Delta \lambda, \Delta p) = \sum_x \|T(x) + \sum_i \lambda_i A_i - I(W(x, p)) + (\nabla T + \sum_i \lambda_i \nabla A_i) \frac{\partial W}{\partial p} \Delta p \sum_i A_i \Delta \lambda_i\|^2 \tag{19}$$

Considering the composed parameters vector $q = \begin{pmatrix} p \\ \lambda \end{pmatrix}$ and $\Delta q = \begin{pmatrix} \Delta p \\ \Delta \lambda \end{pmatrix}$, the optimal can be obtained by

$$\Delta q = -H_{sic}^{-1} \sum_x J_{sic}^T(x) e_{sic}(x),$$

where:

- $e_{sic}(x) = T(x) + \sum_i \lambda_i A_i - I(W(x, p))$
- $J_{sic}(x) = ((\nabla T + \sum_i \lambda_i \nabla A_i) \frac{\partial W}{\partial p_1}, \dots, (\nabla T + \sum_i \lambda_i \nabla A_i) \frac{\partial W}{\partial p_n}, A_1(x), \dots, A_m(x))$
- $H_{sic} = \sum_x J_{sic}^T J_{sic}$

3.3 Projected Out Optimization

The last method we analyze is the project out algorithm, based in a idea from Hager and Behumeur [6] to decompose the optimization in two steps. It was reformulated later in [1] using the inverse compositional scheme.

Representing the linear subspace spanned by the vectors A_i by $span(A_i)$ and its orthogonal complement by $span(A_i)^\perp$, the equation (13) can be rewritten as:

$$\begin{aligned} & \|T(x) + \sum_i \lambda_i A_i(x) - I(W(x, p))\|_{span(A_i)}^2 \\ & + \|T(x) + \sum_i \lambda_i A_i(x) - I(W(x, p))\|_{span(A_i)^\perp}^2 \end{aligned} \tag{20}$$

In previous expression, two terms appear:

- The first term is a non-linear optimization with respect to the warp parameters, but performed in a subspace in which the appearance variation can be ignored (projected out).
- The second step is a closed form linear optimization with respect to the appearance parameters.

The second term can be simplified taking into account that the norm in the second term only considers the component of the vector in the orthonormal complement of $span(A_i)$. Then, component in $span(A_i)$ can be dropped. Then, the resulting expression for minimization is as follows:

$$\begin{aligned} & \|T(x) + \sum_i \lambda_i A_i(x) - I(W(x, p))\|_{span(A_i)}^2 \\ & + \|T(x) - I(W(x, p))\|_{span(A_i)^\perp}^2 \end{aligned} \tag{21}$$

The second of these two terms does not depend upon λ . For any p , the minimum value of the first term is always exactly 0 because the term $\sum_i \lambda_i A_i(x)$ can represent any vector in $span(A_i)$. As a result, the simultaneous minimum over both p and λ can be found sequentially by first minimizing the second term with respect to p alone, and then treating the optimal value of p as a constant to minimize the first term with respect to λ .

The optimal pose obtained from the second term can be expressed as:

$$\Delta p = -H_{po}^{-1} \sum_x J_{po}^T(x) e_{po}(x)$$

where:

- $e_{po}(x) = T(x) - I(W(x, p))$
- $J_{po}(x) = \nabla T \frac{\partial W}{\partial p}(x, 0) - \sum_i \sum_y A_i(y) \nabla T \frac{\partial W}{\partial p}(y, 0) A_i(x)$
- $H_{po}(x) = \sum_x J_{po}^T(x) J_{po}(x)$

Computation of J_{p_o} is carried out by projecting $\nabla T \frac{\partial W}{\partial p}(x, 0)$ vectors into $span(A_i)^\perp$ by removing the component in the direction of A_i , for $i = 1, \dots, m$ in turns. See [1] for more details about computing the $span(A_i)$.

The optimal appearance parameters are given from the first term as:

$$\lambda = (A^T * A)^{-1} * \sum_x A^T e_{p_o}(x)$$

4 Second Order SSD Linear Appearance Models Tracking

In this section the second order formulation is applied to the optimization algorithms of the previous section.

4.1 Independent Optimization of Pose and Appearance

The approximation (10) must be used instead of the first order approximation in (15), originating these new equations for the Δp parameters computing:

$$\Delta p = H_{iic-esm}^{-1} \sum_x J_{iic-esm}^T e_{iic-esm}(x),$$

where:

- $e_{iic-esm}(x) = T(x) - I(W(x, p)) - \sum_i \lambda_i A_i(W(x, p))$
- $J_{iic-esm} = \frac{1}{2}(\nabla_x T + \nabla_x I(W(x, p))) \frac{\partial W}{\partial p}$
- $H_{iic-esm} = \sum_x J_{iic-esm}^T J_{iic-esm}(x)$

As result, the Jacobian $J_{iic-esm}$ (and $H_{iic-esm}$) depends on p parameters, so it has to be computed at each iteration.

4.2 Simultaneous Pose and Appearance Optimization

In (17), the approximation (10) can be performed, giving the following solution

$$\Delta q = -H_{sic-esm}^{-1} \sum_x J_{sic-esm}^T e_{sic-esm}(x),$$

with:

- $e_{sic-esm}(x) = T(x) + \sum_i \lambda_i A_i - I(W(x, p))$
- $J_{sic-esm}(x) = (\frac{1}{2}(\nabla_x T + \nabla_x I(W(x, p)))) + \sum_i \lambda_i \nabla A_i \frac{\partial W}{\partial p_1}, \dots, (\frac{1}{2}\nabla T + \nabla I(W(x, p)) + \sum_i \lambda_i \nabla A_i) \frac{\partial W}{\partial p_n}, A_1(x), \dots, A_m(x)$
- $H_{sic-esm} = \sum_x J_{sic-esm}^T J_{sic-esm}$

Please note that the Jacobian $J_{sic-esm}(x)$ also has terms coming from the linear approximation of $A(W(x, \Delta p))$. As no "second order" approximation is implemented for that expression, the whole algorithm does not reach a "second order" convergence but a lower one.

4.3 Projected Out Optimization

Again, using the approximation (10) in the second term of (21), it can be written:

$$\Delta p = -H_{po-esm}^{-1} \sum_x J_{po-esm}^T(x) e_{po-esm}(x),$$

with:

$$\begin{aligned} - e_{po-esm}(x) &= T(x) - I(W(x, p)) \\ - J_{po-esm}(x) &= \frac{1}{2}(\nabla T + \nabla I(W(x, p))) \frac{\partial W}{\partial p}(x, 0) \\ &\quad - \sum_i \sum_y A_i(y) \frac{1}{2}(\nabla T + \nabla I(W(x, p))) \frac{\partial W}{\partial p}(y, 0) A_i(x) \\ - H_{po-esm}(x) &= \sum_x J_{po-esm}^T(x) J_{po-esm}(x) \end{aligned}$$

5 Experimental Results

In this section the empirical validation of the proposed algorithms is performed by comparing the obtained results. In the experiments, in order to have a ground truth, the algorithms have been tested by warping a static image; homographies (eight parameters warps) were generated by randomly perturbing the four corners of the target region previously aligned with its instance in the tracked image. Then the different tracking algorithms were executed trying to recover the initial pose. Two magnitudes were measured: the *average frequency of convergence* and the *average rate of convergence*. We perform as much experiments as needed to let all the algorithms converge at least 10 times, and 20 iterations per experiment were used. We have used several sets of images to check the consistency of the results across different tests.¹

5.1 Image Tracking Algorithms Without Appearance Variation

The first experiment compares the results obtained with two image tracking algorithm without appearance variation: inverse compositional alignment (*ic*) and our implementation of second order inverse compositional alignment (*esm ic*).

Figure 1 plots the convergence rate obtained from both algorithms using three different magnitudes for point displacement. As it can be seen, the speed of convergence of the *esm* algorithm is higher; that means we can use less iterations to converge, saving computational cost. It also shows the mean frequency of convergence (% over the total number of tests) showing how the frequency of convergence of the *inverse compositional* algorithm decays quicker than the *esm*.

5.2 Lineal Models Tracking Algorithms

Several experiments were carried out to test the behavior of the proposed second order minimization algorithms with respect to their linear order version. It has

¹ More experimental results and tracking video sequences can be obtained at <http://www.ac.uma.es/~jgmora/acivs06>

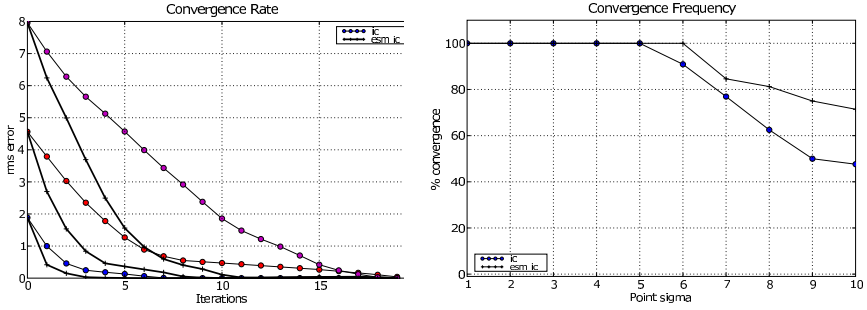


Fig. 1. SSD Image Tracking Results

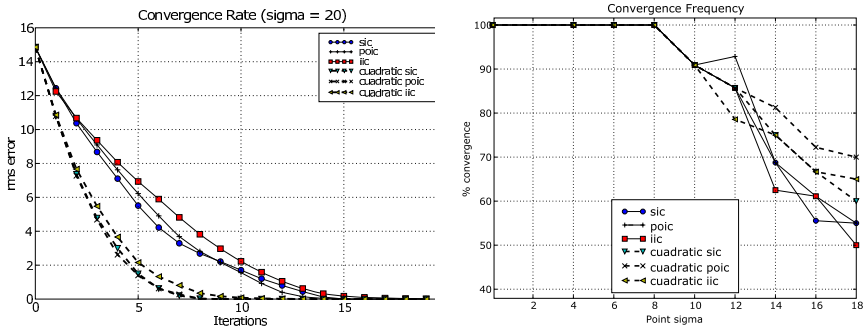


Fig. 2. SSD Linear Model Tracking Convergence rate

been employed a linear basis obtained from three images from different faces and PCA analysis was applied. A maximum spatial error of one pixel is given as criterion for convergence. Figure 2 represents the convergence rate corresponding to the point perturbation ($\sigma = 20$), showing a better behavior of the second order minimization algorithm that require a short number of iterations for convergence. It also includes the convergence frequency for linear models tracking and it illustrates that quadratic algorithms converge, in general, in more occasions.

6 Conclusions

In this paper an analysis of the second order minimization technique has been carried out. As a result a family of algorithms for linear model tracking have been developed following the inverse compositional approach. These techniques have been successfully employed in the tracking of changing appearance targets, showing a better behavior than traditional first order based approximations. This is illustrated in the experimental results, with high convergence rates and improved convergence domain.

As seen in the problem formulation, second order approach results in higher complexity at each iteration because some steps cannot be computed offline as in first order algorithms. However, as pointed out in [2], the quicker convergence of the second order minimization algorithm (in terms of the required steps to converge) can be used to compensate the additional computational cost of each iteration due to the new elements computed. In addition, in the linear model tracking case for algorithms as Simultaneous Inverse Compositional alignment is necessary to recalculate different expressions (section 3.2) even for the first order inverse compositional approach because of its dependency from the appearance parameters computed at each iteration. Thus, the new calculations carried out by the second order approach do not have a big impact on the overall computation. In this subject, our current aim is the development of an efficient implementation of the algorithms to achieve real-time execution similar to that obtained from some linear approximations with the improved behaviour (convergence frequency) of the second order approach. At this moment we are evaluating the use of GPUs exploiting the parallelism that appears in the operations.

Future works will also involve a more in deep study of other related parameters as robustness to corrupting noise and occlusions. We will also study the applicability of these techniques to the shape deformation handling.

References

1. Baker, S., Gross, R., Matthews, I.: Lucas-Kanade 20 Years On: A Unifying Framework: Part 3 Technical Report CMU-RI-TR-03-35. (2004)
2. Benhimane, S., Malis, E.: Real-time image-based tracking of planes using efficient second-order minimization. Proc. IROS04 (2004)
3. Baker, S., Matthews, I.: Lucas-Kanade 20 years on: A unifying framework: Part 1 Int. Journal of Computer Vision **56(3)** (2004) 221–255
4. Baker, S., Matthews I.: Equivalence and efficiency of image alignment algorithms. Proc. of International Conference on Computer Vision and Pattern Recognition (2001) 1090–1097
5. Lucas, B.D., Kanade, T.: An iterative image registration technique with application to stereo vision. IJCAI81 (1981) 674–679
6. Hager, G., Belhumeur, P. Belhumeur: Efficient region tracking with parametric models of geometry and illumination. IEEE Transactions on Pattern Analysis and Machine Intelligence (1998) 1025–1039
7. Black, M., Jepson, A.: Eigen-tracking: Robust matching and tracking of articulated objects using a view-based representation. International Journal of Computer Vision **26(1)** (1998) 63–84
8. Buenaposada, J., Munoz, E., Baumela, L.: Efficient appearance-based tracking. IEEE Workshop on Articulated and Nonrigid Motion (2004)
9. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active appearance models IEEE Transactions on Pattern Analysis and Machine Intelligence (2001) 681–685

Visibility of Point Clouds and Mapping of Unknown Environments

Yanina Landa¹, Richard Tsai², and Li-Tien Cheng³

¹ Department of Mathematics, University of California, Los Angeles, CA 90095
ylanda@math.ucla.edu

² Department of Mathematics, The University of Texas at Austin, TX 78704
ytsai@math.utexas.edu

³ Department of Mathematics, University of California, San Diego, CA 92093
lcheng@math.ucsd.edu

Abstract. We present an algorithm for interpolating the visible portions of a point cloud that are sampled from opaque objects in the environment. Our algorithm projects point clouds onto a sphere centered at the observing locations and performs essentially non-oscillatory (ENO) interpolation to the projected data. Curvatures of the occluding objects can be approximated and used in many ways. We show how this algorithm can be incorporated into novel algorithms for mapping an unknown environment.

1 Visibility

The problem of visibility involves the determination of regions in space visible to a given observer when obstacles to that sight are present. When the observer is replaced by a light source in the simplified geometrical optics setting with perfectly absorbing boundary condition at the obstacles, the problem translates to that of finding illuminated regions. In this regard, the visibility problem is highly related to the high frequency wave propagation problems and is needed in many computational high frequency wave approaches [2]. We will interchange the term visibility with illumination, and occlusion with shadow freely in this paper.

In visualization, visibility information can be used to make complicated rendering processing more efficient by skipping over occlusion. In robotics mission planning, achieving certain visibility objectives may be part of the mission. Video camera surveillance design is one such example.

Visibility problems have also been studied by geometers. For example, H. Wentz asked if connectedness of the on surface shadow is sufficient to imply convexity of the occluding surface [4].

In general, one may consider the following classes of visibility problems:

1. Given occluders, construct shadow volume and its boundary;
2. Given a projection of visible regions, construct the occluders;
3. Find location(s) that maximize visibility using certain predefined metric.

In many visualization applications, (1) is solved by projecting triangles. Wentz's question can be viewed as in category (2). Problems related to surveillance is

related to (2). We will present an algorithm for a problem related to both (1), (2), and (3).

1.1 Representations of Visibility

Today computational geometry and combinatorics are the primary tools to solve visibility problems [5][18],[3]. The combinatorial approach is mainly concerned with defining visibility on polygons and more general planar environments with special structure. All the results are based on an underlying assumption of straight lines of sight. The simplified representation of the environment is a major limitation of this methodology. Furthermore, the extension of these algorithms to three dimensional problems may be extremely complicated.

Our goal is to define such a representation of visibility as to be able to solve the problems considered in computational geometry [5] on general environments in two or three dimensions, independent of the integral field defining the lines of sight, utilizing minimum information about the environment.

One attempt was to introduce the level set representation of the occluding objects and the visibility function, defined in [16]. While this algorithm can be applied to general types of environment, easily extended to three dimensions, and curved lines of sight, it requires a priori knowledge of the occluding objects to construct the level set representation of the environment. This information may not be available in some important real life applications, e.g. navigation in an unknown environment, or if the occluding objects are represented by open surfaces.

Another method for visibility representation was introduced by LaValle et al in [6], [14]. This is a rather minimal framework based on detecting discontinuities in depth information (called gaps) and their topological changes in time (referred to as gap critical events). The “visible” environment is represented by a circle centered at the vantage point, with gaps marked on the circumference in the order of their appearance to the observer. Note that no distance or angular information is provided. As with most combinatorial approaches, LaValle’s method works only on regions having special geometries.

In [19], an algorithm extracting planar information from point clouds is introduced and used in mapping outdoor environment. In [11], depth to the occluders is estimated by a trinocular stereo vision system and is then combined with a predetermined “potential” function so that a robot can moved to the desired location without crashing into obstacles.

Here we introduce a new model which, similarly to the level set representation, can handle complicated geometries and curved lines of sight. In contrast to LaValle’s representation, we utilize distance and angular information, which, in practice, can be easily provided by the sensor.

2 Visibility Interpolation and Dynamics

Assume we have a set of points P that are “uniformly” sampled from the occluding surfaces. In practice this data could be obtained from sensors such as LIDAR

or even from triangulated surfaces (here P would be the set of vertices). Given a vantage point, our algorithm would produce a subset of visible data points and a piecewise polynomial interpolation of the visible portions of the surfaces. Unlike the level set representation [16], our algorithm can handle open surfaces and does not require a priori knowledge of occluding surfaces to construct visibility.

2.1 Basic Formulation

Let us begin by introducing some notations. Let x_0 denote the vantage point (always assume x_0 outside of the objects). Consider $\Omega \subset \mathbb{R}^d$ ($d = 2, 3$) – a set of objects in question, $\Gamma = \partial\Omega$, and $\Gamma_{x_0}^*$ – visible portion of Γ with respect to x_0 . Denote by ϕ the signed distance function to Γ . We define the view direction from x_0 to x by $\nu(x_0, x) := (x - x_0) / |x - x_0|$. For any two points in space x_1 and x_2 , we say that $x_1 \leq x_2$ (x_1 is “before” x_2) if $\nu(x_0, x_1) = \nu(x_0, x_2)$ and $|x_1 - x_0| \leq |x_2 - x_0|$. Also, a point $y \in \Gamma$ is called a horizon point if and only if $\nu(x_0, y) \cdot n(y) = 0$, where $n(y)$ is the outer normal of Γ at y . Lastly, a point $y \in \Gamma$ is called a cast horizon point if and only if there is a point y^* such that $y^* \leq y$ and y^* is a horizon point.

Observe that the visibility status of points sharing the same radial direction with respect to the vantage point satisfies a causality condition. That is, if x_1 is occluded and $x_1 \leq x_2$, then x_2 is also occluded. We set

$$\rho_{x_0}(p) := \begin{cases} \min_{x \in \bar{\Omega}} \{|x - x_0| : \nu(x_0, x) = p\}, & \text{if exists} \\ \infty, & \text{otherwise} \end{cases} \tag{1}$$

Define the visibility indicator $\Theta(x, x_0) := \rho(\nu(x, x_0)) - |x - x_0|$ such that $\{\Theta \geq 0\}$ is the set of visible regions and $\{\Theta < 0\}$ is the set of occluded regions. See Fig. 1 for an example.

Assume, in addition, that the sampling of points is “uniform”. That is, we can find an $\epsilon > 0$, such that ϵ -balls centered at each sampled point on Γ connect the connected components and do not connect disconnected components of Γ .

Let $P \subset \mathbb{R}^d$ be the sampled data set. Enumerate all the points $y_i \in P$. Define the projection operator $\pi_{x_0} : \mathbb{R}^d \mapsto S^{d-1}$, mapping a point onto the unit sphere centered at x_0 . Then we can construct the following piecewise constant approximation to the surface on a sphere:

$$\tilde{\rho}_{x_0}(z) = \min(\rho_{x_0}(z), |x_0 - y_i|), \text{ for every } z \in \pi_{x_0}B(y_i, \epsilon). \tag{2}$$

In addition we can define an auxiliary function $R_{x_0} : S^{d-1} \mapsto P$, which records $\tilde{P} \subset P$ – a subset of all points in P visible from x_0 :

$$R_{x_0}(z) := \begin{cases} y_i, & \text{if } \rho_{x_0}(z) > |x_0 - y_i| \\ \text{value unchanged,} & \text{otherwise} \end{cases} \tag{3}$$

In case the surface normals are available for each data point, we can use ellipse instead of a ball in the above construction. In [12], a similar projection approach is proposed for rendering purposes.

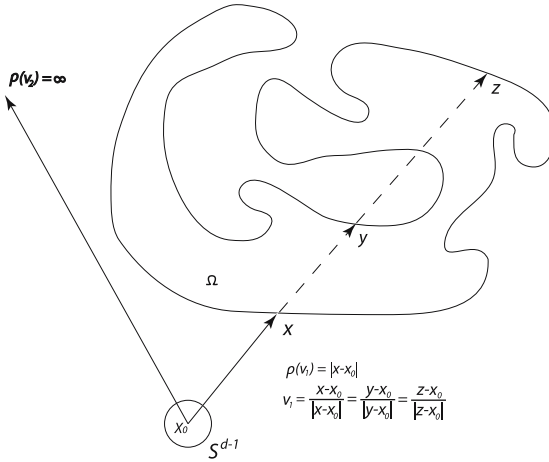


Fig. 1. Demonstration of visibility

2.2 Smoother Reconstruction by ENO Interpolation

Note that analytically the visibility function ρ is piecewise continuous with jumps corresponding to the locations of horizons. Smoothness of ρ in each of its continuous pieces relates to the smoothness of the corresponding visible part of Γ , i.e. $\Gamma_{x_0}^*$. In the previous section we obtained a piecewise constant approximation $\tilde{\rho}_{x_0}$ of the visibility function and recorded an auxiliary function R_{x_0} which keeps track of the visible data points serving as “originators” of the constant values of $\tilde{\rho}_{x_0}$. We will use R_{x_0} to construct a piecewise polynomial approximation ρ_{int} to the visibility function which would preserve the jumps. ENO (Essentially Non-Oscillatory) interpolation introduced by Harten et al [7] is used to compute such a ρ_{int} .

For example, consider a two dimensional reconstruction on S^1 . First, parameterize S^1 by angles $\theta \in [-\pi, \pi)$. Then sort the visible points $p_i \in \tilde{P}$ in the increasing order of the angle they form with respect to the vantage point: $\rho_{x_0}^{-1}(p_i) = \arg(p_i - x_0)$. To construct a piecewise linear interpolation $\rho_{x_0}^{ENO(1)}$ use the values of $\tilde{\rho}_{x_0}(\theta)$, where $\theta \in I[\tilde{\rho}^{-1}(p_i), \tilde{\rho}^{-1}(p_{i+1})]$. Similarly, we can obtain $\rho_{x_0}^{ENO(p)}$ – a piecewise p -th order interpolation. See Fig.2 for an example.

ENO interpolation can be applied in two steps to compute an approximation on S^2 for organized data clouds. Let θ_1 and θ_2 parameterize S^2 . We first ENO-interpolate $\rho_{x_0}^{(0)}(\cdot, \theta_2)$ in the θ_1 direction to obtain $\rho_{x_0}^{ENO(p,*)}$. Then use $\rho_{x_0}^{(0)}(\theta_1, \cdot)$ and $\rho_{x_0}^{ENO(p,*)}$ to interpolate in θ_2 direction to obtain $\rho_{x_0}^{ENO(p,q)}$. Figures 3 and 4 are examples in three dimensions.

We shall use the piecewise p -th order approximation $\rho_{x_0}^{ENO(p)}$ to compute derivatives on the occluding surfaces (away from the edges) and easily extract various geometric quantities.

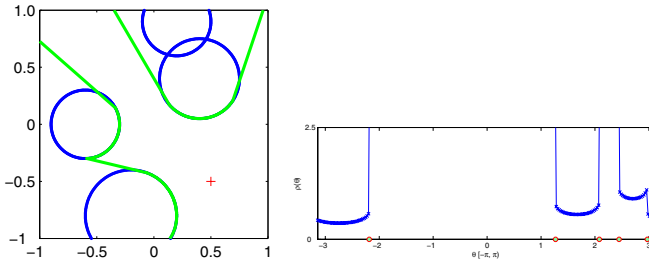


Fig. 2. Points visible from $(0.5, -0.5)$, corresponding visibility function $\rho(\theta)$, and the edges (horizon points)

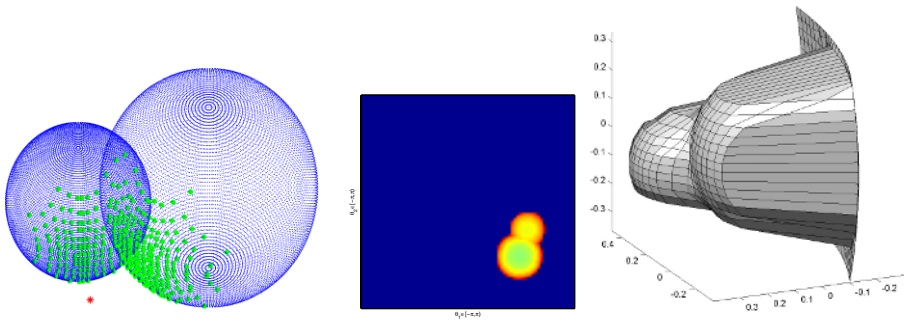


Fig. 3. Visible points from a vantage point marked by red star, corresponding visibility function $\rho(\theta_1, \theta_2)$, and the reconstructed visible surface

2.3 Curved Lines of Sight

To demonstrate the flexibility of our formulation, consider the case when the lines of sight are no longer straight. Then we can not use the relation $\nu(x, x_0) = (x - x_0)/|x - x_0|$ in the definition of the visibility function (1). As in [16], we consider instead the flow lines connecting x_0 to the data points $p \in P$. The construction of the visibility function is done as follows. First, we construct the distance function φ on the whole domain D by solving the eikonal equation

$$|\nabla\varphi(x)| = r(x), \text{ in } D, \varphi(x_0) = 0, \tag{4}$$

where $r(x) > 0$ is the variable index of refraction. We use the fast sweeping technique from [17] to solve (4). To determine the polar coordinates $(\theta, \rho(\theta))$ corresponding to the point \mathbf{p} on the occluding surface we then solve

$$\begin{aligned} \frac{\partial x}{\partial t} &= -\nabla\varphi(x), \\ x|_{t=0} &= p, \end{aligned} \tag{5}$$

to trace point p back to x_0 along the line of sight connecting them. Then θ is the angle made by $\nabla\varphi$ at x_0 , and $\rho(\theta) = \varphi(p)$. The visibility function can be constructed using the causality condition with respect to φ . See Figure 5.

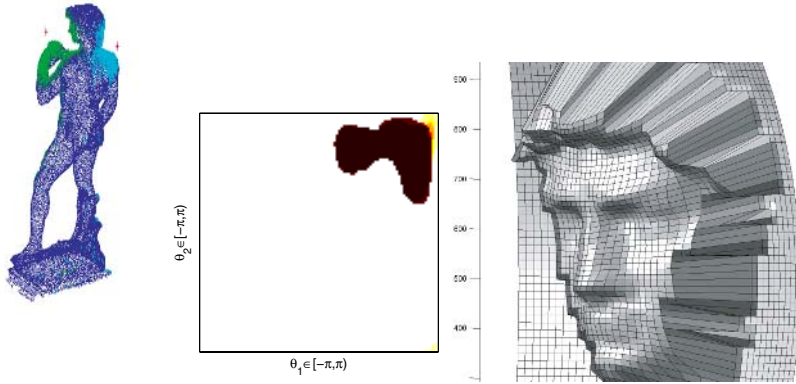


Fig. 4. Visible points on Michelangelo’s statue [10] from two vantage points, one of the corresponding visibility maps $\rho(\theta_1, \theta_2)$, and a reconstruction of portion of the visible surface

Such computations may be useful when determining visibility in regions with variable refraction such as water or fog, or in anisotropic medium (in this case, one needs to solve more general Hamilton-Jacobi equations as considered in [16]).

2.4 Dynamics

When the lines of sight are straight, we can derive how the visibility changes along with a moving vantage point x_0 . In two dimensions let us consider a coordinate system centered at x_0 with the visible portions of the occluding surfaces parameterized by polar coordinates. A point z on the occluder is visible from x_0 . Assume the observer moves with the velocity $v = (v_1, v_2)$. The value of the visibility function is $\rho_{x_0}(\theta) = |z - x_0|$. Suppose during the period of time Δt the observer has moved to a new location $x_0 + v\Delta t$. The corresponding value of the visibility function is $\tilde{\rho}_{x_0+v\Delta t}(\tilde{\theta}) = |z - (x_0 + v\Delta t)|$. The angle between the velocity vector v and the x -axis is $\phi = \tan^{-1} \frac{v_2}{v_1}$. The angle between $z - x_0$ and the velocity vector v is ψ . Then, the angle between $z - x_0$ and the x -axis is $\theta = \phi + \psi$.

We can obtain the following expressions:

$$\frac{d\theta}{dt} = |v| \sin \psi, \tag{6}$$

$$\frac{d}{dt} \left(\rho(\theta(t), t) \right) = \rho_t + \rho_\theta \theta_t = \frac{d}{dt} |x_0(t) - z|. \tag{7}$$

Now we can put (6) and (7) together to get

$$\rho_t + |v(t)| \sin \psi \rho_\theta = v(t) \cdot \begin{pmatrix} \cos \theta \\ \sin \theta \end{pmatrix}. \tag{8}$$

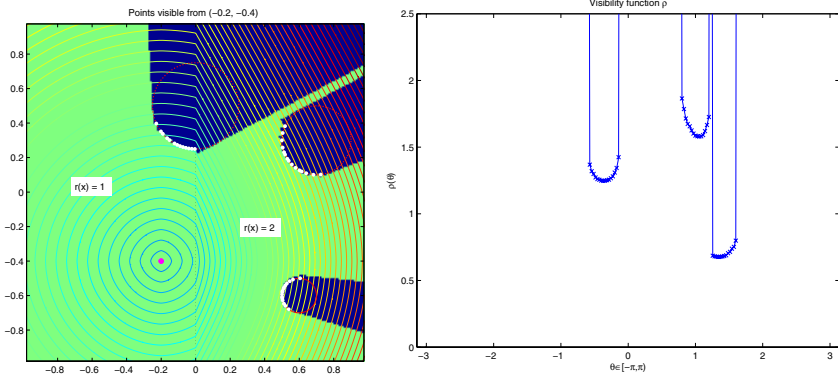


Fig. 5. Left: non-straight lines of sight; Right: corresponding visibility function $\rho(\theta)$

Now let us consider the motion of horizon points e_1 and e_2 . Note that $(e_i - x_0) \cdot n_{e_i} = 0$, where n_{e_i} is the outer unit normal to the occluding surface at the point e_i for $i = 1, 2$. That is, $e_i - x_0$ is tangent to the occluding surface at the horizon point. Without loss of generality, in all future computations we will consider just e_1 .

In the coordinate system centered at x_0 , $\theta = \phi + \psi$ is the angle between $e_1 - x_0$ and the x -axis. The value of the visibility function is $\rho_{x_0}(\theta) = |e_1 - x_0|$. Now suppose the observer moves to a new position $x_0 + v\Delta t$, moving with the velocity $v = (v_1, v_2)$. For this new location, the position of the edge has changed to \tilde{e}_1 and the corresponding value of the visibility function is $\tilde{\rho}_{x_0 + v\Delta t}(\tilde{\theta}) = |\tilde{e}_1 - (x_0 + v\Delta t)|$. Here $\tilde{\theta} = \phi + \tilde{\psi}$ is the angle between $\tilde{e}_1 - (x_0 + v\Delta t)$ and the x -axis in the coordinate system centered at $x_0 + v\Delta t$. Our goal is to find the change in the position of horizon, i.e. $\frac{d}{dt}e_1$.

First, note that the curvature of the occluding surface at the point $(\rho(\theta), \theta)$ is given by

$$\kappa = \frac{\rho^2 + 2\rho_\theta^2 - \rho\rho_{\theta\theta}}{(\rho^2 + \rho_\theta^2)^{\frac{3}{2}}}. \tag{9}$$

Also, since $e_1 - x_0$ is tangent to the occluder at e_1 , we obtain

$$\begin{aligned} n^\perp(e_1) &= \frac{e_1 - x_0}{|e_1 - x_0|} \\ n(e_1) &= \left(n^\perp(e_1)\right)^\perp = \left(\frac{e_1 - x_0}{|e_1 - x_0|}\right)^\perp. \end{aligned} \tag{10}$$

Now we can plug in the above into the formula for horizon dynamics from [16] to get

$$\frac{de_1}{dt} = \frac{1}{\kappa} \frac{v \cdot n(e_1)}{|e_1 - x_0|} n^\perp(e_1), \tag{11}$$

or, using the fact that $v \cdot n(e_1) = |v| \cos(\psi + \frac{\pi}{2})$,

$$\frac{de_1}{dt} = \frac{|v| \cos(\psi + \frac{\pi}{2})}{\kappa \rho^2} (e_1 - x_0). \tag{12}$$

Remember that in all of the above $\psi = \theta - \phi = \theta - \tan^{-1} \frac{v_2}{v_1}$.

Therefore, from (8) and (12) we obtain full description of the change in the visible portion of the occluder with respect to the observer’s motion.

The corresponding expressions can also be derived in three dimensions, see [16].

3 Applications of Visibility Interpolation to Navigation Problems

Let us consider the application of visibility to navigation in an unknown environment, for example exploring the environment, object finding, and pursuit-evasion. LaValle et al have addressed these problems in [6], [14], [15], [13], [9]. Their algorithms only work on polygonal domains or curved regions whose boundary may be represented as a set of solutions to an implicit polynomial equation of the form $f(x_1, x_2) = 0$ (see [9]). Our algorithms work on general types of environments using point cloud data that is either presampled or sampled in action by some hardware.

3.1 Problem: Seeing the Whole Environment

Here we consider the problem of exploring the unknown bounded region with obstacles. The objective is to map the whole environment. We set the following restrictions on the path traveled by the observer:

1. The path should be continuous and consist of discrete steps;
2. The number of steps should be finite;
3. The total distance traveled must be finite.

These restrictions ensure that the algorithm would be practical in real life applications. Consider first simple, but non-practical examples of navigation in a bounded region with a single occlusion in shape of a circle. One strategy to explore the environment around the occlusion would be to approach the circle’s boundary and travel along it until we return to the same point. This strategy does not satisfy our restrictions since it would require an infinite number of steps to travel along the boundary. Another strategy would be to proceed to infinity to see half a circle at once, then jump to infinity at the opposite side of the circle to see the other half. Such a strategy does not satisfy our restrictions either, since the path would be infinite and not continuous.

Our algorithm was inspired by LaValle et al. In this method the observer randomly chooses a gap marked on the visibility plot and approaches it. The visibility map is then updated and the process is repeated until the whole region

is explored. Critical events such as appearance and disappearance of gaps are tracked by the dynamic data structure. Since the visibility map has no distance or angular information, the algorithm is not optimal with respect to the total distance traveled. In particular, if this algorithm is applied to cases that contain fine polygonalization of curved objects, the computational cost of this algorithm may become too large.

Our visibility representation includes distance and angular information, and our algorithm is designed with the consideration of handling basic smooth geometries. In essence, the visibility of any bounding disk of a convex object guarantees the visibility of that object. If a set of separated, non-overlapping bounding disks exists for a collection of disjoint convex objects, we may consider the visibility problem of each convex object independently. Furthermore, if a non-convex object can be decomposed by the set difference of a finite number of convex sets, then one can treat it “almost” like a convex object. We see that the signed curvature, a notion of local convexity, is rather essential in applying the above arguments. We obtained formulas for the upper bounds of the number of observing locations in these situations as functions of the sign changes in the curvatures as well as the number of disconnected components, and would report our finding in a forthcoming paper.

ALGORITHM 1

1. For the given x_0 outside the occluding objects; construct the visibility function $\rho(\theta)$;
2. Find all the edges on the $(\theta, \rho(\theta))$ map and proceed to the nearest edge;
3. Find edges;

If no edges are found, we are on the boundary of an obstacle at the horizon point. Thus we need to “overshoot” x_0 along the tangent line to see where to proceed next. We choose the following overshooting step size

$$r = \lambda \tan\left(\frac{\pi}{3}\right) \frac{1}{\kappa}, \quad (13)$$

where κ is the curvature of an edge defined by (9) and λ is a parameter. This way we have a minimal number of steps to travel around the obstacle, e.g. for a circle, $r = 1/2$ the side of the equilateral triangle enclosing the circle. In case $\kappa = 0$ we shall shift the position by a small amount to see the next edge.

If the edges are found, move x_0 to the nearest edge. Store the unexplored edges in a list;

4. Finish when the change in total visible area is less than the desired tolerance and all the edges are “removed” from the list. Otherwise go to 1 with the current location of x_0 .

Figure 6 illustrates the steps of the above algorithm with one and two circles as obstacles. The pink arcs correspond to the portions of the circles that are reconstructed. Figure 7 depicts final paths for different test cases. As one can

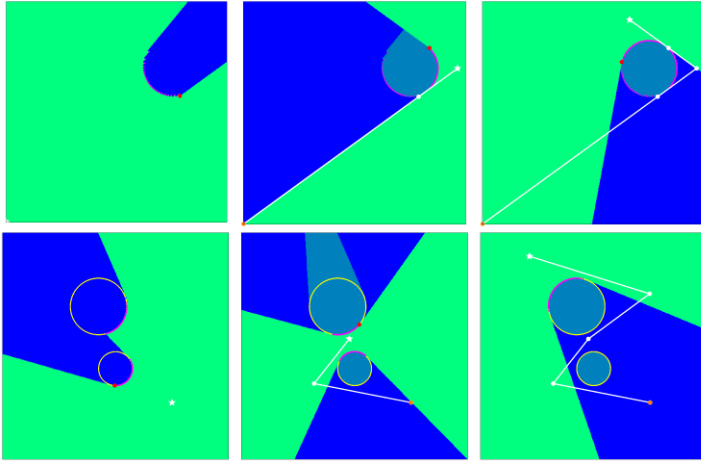


Fig. 6. Steps of the exploration algorithm with one and two circles as obstacles

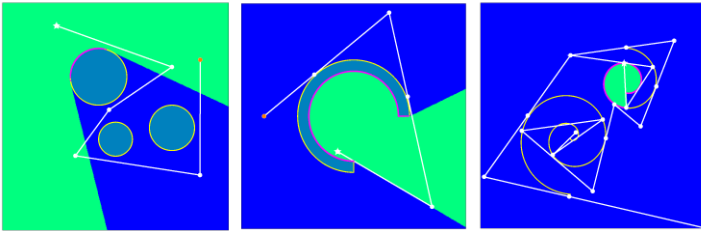


Fig. 7. Full paths for different obstacles

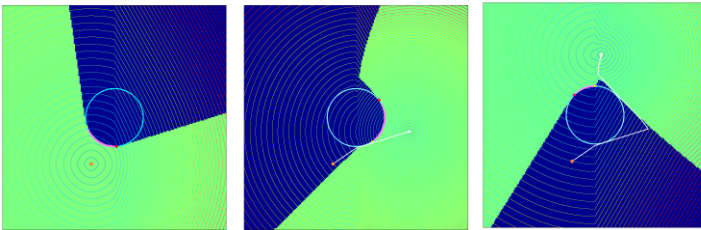


Fig. 8. Full path in curved rays

see from the examples, the algorithm handles both convex and concave obstacles. The algorithm always converges, however it does not provide the desired optimality with respect to the total distance traveled.

We remark that ALGORITHM 1 can be applied to the curved lines of sight cases with following modifications. Since the curvature of the occluding surfaces cannot be recovered from the visibility function ρ , the overshoot step-size must

be defined by the user in step 3. To proceed further from the edge we follow the line of sight passing through this edge by solving

$$\frac{dx}{dt} = \nabla\phi(x), x(0) = x_e, \quad (14)$$

where x_e is the position of the edge. Consider Fig.8 for a sample step-by-step path.

4 Conclusion

We present an essentially non-oscillatory algorithm for interpolating point cloud visibility information in polar coordinates. This algorithm is capable of approximating higher order derivatives of the surface so that curvatures can be computed. We also present a new path planning algorithm using our point cloud visibility interpolation. Our future work lies in optimizing the above algorithm. We desire a better performance with respect to the distance traveled and/or the number of steps.

Acknowledgement

Landa's research is supported by ONR MURI Grant N00014-02-1-0720, Tsai's research is supported by NSF DMS-0513394, and Cheng's research is supported by Alfred P. Sloan Fellowship and NSF Grant 0511766.

References

1. A. Atle and B. Engquist, "On surface radiation conditions for high frequency wave scattering", preprint, (2006).
2. O. Bruno, C.A. Geuzaine, J.A. Monro Jr., and F. Reitich, "Prescribed error tolerances within fixed computational times for scattering problems of arbitrarily high frequency: the convex case", *Philos. Trans. R. Soc. Lond. Ser. A Math. Phys. Eng. Sci.*, **1816**,(2004), 629-645.
3. W.-P. Chin and S. Ntafos, "Shortest watchman routes in simple polygons", *Discrete Comput. Geom.*,**6** (1991), 9-31.
4. M. Ghomi, "Shadows and convexity of surfaces", *Annals of Mathematics*, **155** (2002), 281-293.
5. J.E. Goodman, J.O'Rourke, editors "Handbook of discrete and computational geometry", CRC Press LLC, Boca Raton, FL; Second Edition, April 2004
6. L. Guilamo, B. Tovar, S.M. LaValle, "Pursuit-evasion in an unknown environment using gap navigation graphs" *IEEE International Conference on Robotics and Automation*, (2004), under review.
7. A. Harten, B. Engquist, S. Osher, S.R. Chakravarthy, "Uniformly high order accurate essentially nonoscillatory schemes, III," *Journal of Computational Physics*, **71**, (1987), 231-303.
8. H. Jin, A. Yezzi, H.-H. Tsai, L. T. Cheng and S. Soatto, "Estimation of 3D surface shape and smooth radiance from 2D images; a level set approach", To appear, *Journal of Scientific Computing*, **19**, (2003), 267-292.

9. S.M. LaValle, J. Hinrichsen “Visibility based pursuit-evasion: An extension to curved environments”, *Proc. IEEE International Conference on Robotics and Automation*, (1999), 1677-1682.
10. M. Levoy “The Digital Michelangelo Project”
11. D. Murray and C. Jennings. “Stereo vision based mapping for a mobile robot”, *Proc. IEEE Conf. on Robotics and Automation*, (1997).
12. S. Rusinkiewicz and M. Levoy, “QSplat: A multiresolution point rendering system for large meshes”, *SIGGRAPH*,(2000),343-352
13. S. Sachs, S. Rajko, S.M. LaValle “Visibility based pursuit-evasion in an unknown planar environment”, to appear in *International Journal of Robotics Research*, (2003).
14. B. Tovar, S.M. LaValle, R. Murrieta, “Locally-optimal navigation in multiply-connected environments without geometric maps”, *IEEE/RSJ International Conference on Intelligent Robots and Systems*, (2003).
15. B. Tovar, S.M. LaValle, R. Murrieta, “Optimal navigation and object finding without geometric maps or localization”, *Proc. IEEE/RSJ International Conference on Robotics and Automation*, (2003).
16. Y.-H.R. Tsai, L.-T. Cheng, S. Osher, P. Burchard, G. Sapiro, “Visibility and its dynamics in a PDE based implicit framework” *Journal of Computational Physics*, **199**, (2004), 260-290.
17. Y.-H.R. Tsai, L.-T. Cheng, S. Osher, H.-K. Zhao, “Fast sweeping methods for a class of Hamilton-Jacobi equations”, *SIAM J. Numer. Anal.*, **41**(2), (2003) 673-694.
18. J. Urrutia. “Art gallery and illumination problems”, In J. R. Sack and J. Urrutia, editors, *Handbook of Computational Geometry*, pages 973-1027, 2000.
19. D. F. Wolf, Andrew Howard, and Gaurav S. Sukhatme. “Towards Geometric 3D Mapping of Outdoor Environments Using Mobile Robots”, *IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*,(2005),1258-1263.

Adjustment for Discrepancies Between ALS Data Strips Using a Contour Tree Algorithm

Dongyeob Han, Jaebin Lee, Yongil Kim, and Kiyun Yu

School of Civil, Urban and Geosystem Engineering, Seoul National University,
Shinrim-Dong, Kwanak-Ku, Seoul 151-742, South Korea
{hkyon2, dama77, yik, kiyun}@snu.ac.kr
<http://spins.snu.ac.kr>

Abstract. In adjusting for discrepancies between adjacent airborne laser scanning (ALS) data strips, previous studies generally used conjugate features such as points, lines, and surface objects; however, irrespective of the types of features employed, the adjustment process relies upon the existence of suitable conjugate features within the overlapping area and the ability of the employed method to detect and extract the features. These limitations make the process complex and sometimes limit the applicability of developed methodologies because of a lack of suitable features in overlapping areas. To address these problems, this paper presents a methodology that uses the topological characteristics of the terrain itself, which is represented by a contour tree (CT). This approach provides a robust methodology without the restrictions involved in methods that employ conjugate features. Our method also makes the overall process of adjustment generally applicable and automated.

1 Introduction

Since their introduction, airborne laser scanning (ALS) systems have been adapted to a wide range of application areas such as the creation of digital surface models (DSMs) and orthophoto generation because of the system's ability to quickly acquire 3D terrain coordinates over target areas. Present research related to ALS systems is focused on ways to improve the collection and analysis efficiency of ALS data. Despite recent improvements to the system, there remain noticeable systematic errors in overlapping areas of ALS strips. These errors result from inaccurate calibration of the entire measurement system and the limited accuracy of direct geo-referencing via the global positioning system (GPS) and the inertial measurement unit (IMU), including systematic errors; these errors are generated while the ALS system is flying over multiple overlapping strips to cover the target area [1]. Such systematic errors usually result in less meaningful data and a questionable quality of the final product.

With an increased need to adjust for discrepancies, a series of studies have been undertaken in recent years based on matching methodologies using conjugate features, including points to points matching [2], triangular irregular network (TIN) matching [3], points to surface matching [4], surface objects to surface objects matching [1], [5] and line to line matching methods [6], [7]. In these

studies, the chosen conjugate features are detected and extracted from raw ALS data to enable adjustment; however, the lack of suitable features in overlapping areas sometimes limits the applicability of the developed methodologies. When points, linear features, and flat surfaces are identified from ALS data, they are usually identified from man-made structures. These features are readily detected and extracted from raw ALS data from urban areas, but the process is more difficult for raw ALS data from rural and mountainous areas. Another problem is the automated detection and extraction of conjugate features from raw ALS data; these processes are complicated by the difficulty of feature identification and frequently suffer from expensive computation effort. To overcome these limitations, this paper presents a methodology that uses the topological characteristics of the terrain itself. When two neighboring strips overlap, the local height variation in ALS data within the overlapping area increases if discrepancies exist between the strips; this results in a complex terrain topology. Thus, it should be possible to adjust for the discrepancy by finding the transformation that minimizes the topological complexity, i.e., removes the discrepancy.

We use a contour tree (CT) to represent the topology and measure the topological complexity of the terrain. The CT is a fundamental data structure in scientific visualization, mainly used to capture the topological characteristics of a scalar field that represents data in different application areas such as geographic information systems, medical imaging, or scientific visualization [8]. The CT consists of a finite set V of objects, termed vertices, and a finite set E of objects, termed edges. In general, the vertices represent contour lines or spot points and the edges represent the adjacent relationship of the two vertices. Pairs of vertices that determine an edge are adjacent vertices [9]. For every vertex V_i in a contour tree, we can count the total number of neighboring vertices of V_i . When the number of neighboring vertices is equal to one, the vertex V_i is termed a leaf. Leaves are usually assumed to be isolated, and have a locally extreme value of elevation in the CT structure [8], [10]. Therefore, one can easily imagine that the more leaves within a target area, the more complex the topography. Based on this scheme, the topological complexity is measured using the number of leaves of the CT. For example, Figure 1 shows that the number of leaves of the CT increases with the addition of noise to the MATLAB 'peaks' function. In these simulation data, the number of leaves increased from 50 to 1051 by increasing the ratio of noise to the peak function. As mentioned above, when ALS data strips overlap, discrepancies between the strips act to increase the number of leaves of the CT in the overlapping area. Therefore, we can adjust for the discrepancy between ALS data strips by determining the appropriate transformation function between neighboring strips that minimizes the number of leaves of the CT within the overlapping area. We use this methodology in the present paper to perform the adjustment process without the need for the detection and extraction of conjugate features. This new method makes the overall adjustment process of ALS data strips more generally applicable and automated.

In Section 2, we present an overview of the CT and the methodology used to create it in the current paper, while the algorithm used to adjust for measured

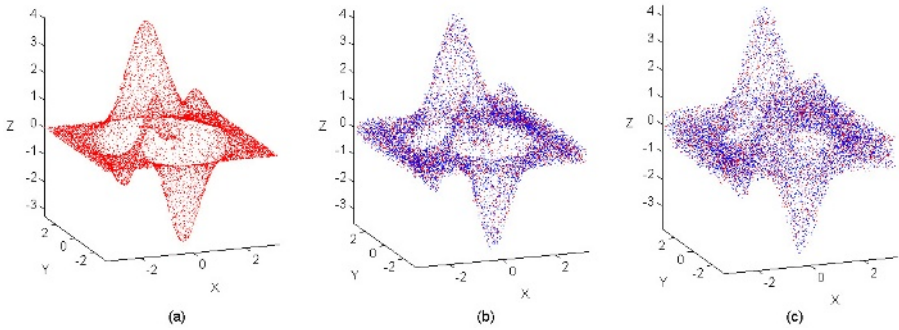


Fig. 1. (a) The original MATLAB 'peaks' function. The magnitude of the noise component added to (b) and (c) is 15% and 30% that of the maximum range of values in the 'peaks' function, respectively. Blue data points represent the added noise.

discrepancies is described in Section 3. In Section 4, we demonstrate the feasibility of this approach via an experiment with real ALS data strips obtained by a state-of-the-art ALS system. Finally, conclusions and future works are discussed in Section 5.

2 Contour Tree

The CT was introduced by [11] as a summary of the elevation of contours on a map (i.e., in 2-D). Since its introduction, the CT has been used in image processing and geographic information systems. The CT is a type of tree structure and a data structure that represents the relationships between connected components of the level sets in a scalar field. The display of the CT provides the user with direct insight into the topology of the field and minimizes the user interaction time that is necessary to "understand" the structure of the data [12]. In particular, points that are usually assumed to be isolated and that have locally extreme values of elevation can easily be detected under the CT structure.

To compute the CT, we use the algorithm proposed by [13], which is an elegant and efficient algorithm for the computation of the CT in any dimension. The algorithm consists of three stages: (i) sorting the vertices in the field, (ii) computing the join tree (JT) and split tree (ST), and (iii) merging the JT with the ST to construct the CT. From a given 3-D point cloud, we can create a mesh M using the Delaunay triangulation, which consists of vertices and edges (see Figure 2). Using the mesh M , the CT is created as described in the following sections.

Sorting the Vertices. The vertices of M are ordered by increasing height values using any standard sorting technique.

Computing the JT and ST. The mesh M can be segmented into different groups that consist of vertices and edges according to the changes in heights between neighboring edges. Computing the JT is the process that determines

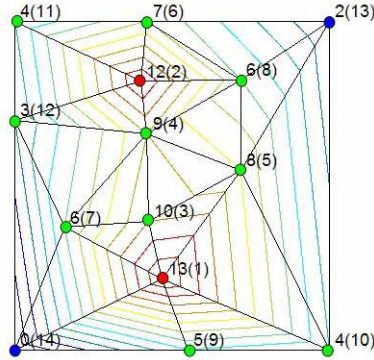


Fig. 2. Representation of the mesh M . The mesh is a bivariate scalar field (terrain) represented as a triangulated irregular network (TIN) with elevation values associated with each vertex. The numbers represent the height value of each vertex, while the numbers in brackets indicate the height order. Critical points are marked with colored disks: local maxima in red and minima in blue.

junction points (see the 4^{th} point in Figure 3) that join groups, while computing the ST is the process that determines junction points (see the 9^{th} point in Figure 3) that split groups from M . When computing the JT, a new group is created if any vertex has a locally maximum value; if not, the group to which the vertex belongs is determined by testing whether the vertex is linked by its edge to other vertices in existing groups. The routine of computing the ST has the same structure as that used in computing the JT, although in reverse. For example, Figure 3 shows the JT and ST created from the mesh M .

Merging the JT and ST to Construct the CT. In the last step of the algorithm, the JT is merged with the ST to construct the CT. The vertices of the JT and ST can be classified as one of three kinds of vertices: leaf vertices, vertices that connect other vertices within the group, and vertices that connect neighboring groups. The upper leaves of the JT and the lower leaves of the ST and their edges are added to the CT. This leaf vertex and its edge are then removed from the JT or ST, and the neighboring vertex of this vertex becomes a new leaf vertex. When all leaf vertices are removed from the JT and ST, the merging process is finished. Using this algorithm, the CT can be computed from a point cloud (see Figure 4).

3 Using the CT to Adjust for Discrepancies

The aim of this paper is to adjust for discrepancies between adjacent ALS data strips. The goal of the adjustment process is to find the appropriate transformation T between strips such that any discrepancy between the transformed strip and reference strip is removed. To use the CT for the adjustment process, we

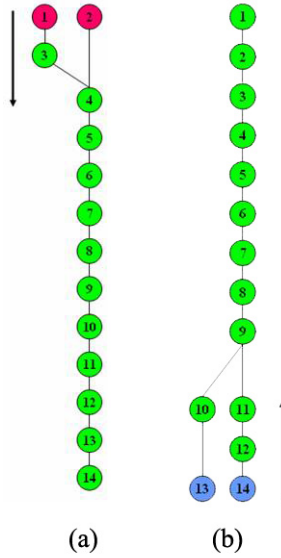


Fig. 3. The correspondence of the join and split tree with the mesh M shown in Figure 2. (a) Join tree. (b) Split tree.

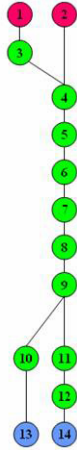


Fig. 4. The correspondence of the contour tree with the mesh M shown in Figure 2

utilize a scheme that dictates that when the discrepancy is perfectly removed, the number of leaves of the CT computed for the overlapping area is minimized. Based on this scheme, the process of determining the appropriate transformation is as follows.

Given two ALS data strips S1 (a test strip) and S2 (a reference strip):

1. First Stage
 - (a) Extract the overlapping area of S1 and S2.
 - (b) Choose the initial spatial transformation T.
 - (c) Determine the range and cell size of parameters of T; these values are dependent on the quality of the approximations of the parameters and the point density of the ALS data.
2. Second Stage
 - (a) Apply the T to S1 and compute the CT and the number of leaves of the CT (N (CT)) of the overlapping area, refining the parameters of T.
 - (b) Find the parameters of T when N (CT) is minimized.
3. Third Stage
 - (a) Decrease the range and cell size of parameters.
 - (b) Repeat Step 2 until the parameters converge to the desired precision.

4 Experimental Results

We applied the developed algorithm to an ALS dataset captured using an OPTECH ALTM 2050 laser scanner at a mean flying altitude of 975 m and a mean point density of 2.24 points/ m^2 . The first and last returns were recorded as well as range and intensity data. According to the sensor and flight specifications, 0.5 m horizontal and 0.15 m vertical accuracies were expected for this dataset. For the target area, the ALS dataset consists of six strips, of which two neighboring strips were used for the present analysis. A total of 1.7 million points were identified from the overlapping area between the two selected strips. To minimize computation time and increase efficiency, the central part of the overlapping area, which includes 616,745 points, was used for our analysis (see Figure 5). A 3-D conformal transformation was chosen as a spatial transformation for adjustment. A 3-D conformal transformation consists of scale, rotation, and translation factors, as apparent in Equation (1) below:

$$\begin{pmatrix} X_r \\ Y_r \\ Z_r \end{pmatrix} = \begin{pmatrix} X_T \\ Y_T \\ Z_T \end{pmatrix} + SR(\omega, \phi, \kappa) \begin{pmatrix} X \\ Y \\ Z \end{pmatrix} \quad (1)$$

where S is the scale factor, $(X_T, Y_T, Z_T)^T$ is the translation vector between the origins of the coordinate systems for each ALS data strip, R is the 3-D orthogonal rotation matrix, $(X, Y, Z)^T$ are the point coordinates in the test strip, and $(X_r, Y_r, Z_r)^T$ are the point coordinates in the reference strip.

After setting up the spatial transformation, we determined the initial parameters of the transformation. The transformation was applied to the test strip, and the CT and the number of leaves of the CT were computed for the overlapping area. Within a predetermined range and cell size of parameters (in this experiment the range and cell size were ± 1 m and 0.2m respectively), seven parameters of transformation were refined sequentially until the number of leaves was minimized. Then, the range and cell size of parameters were reduced by about half

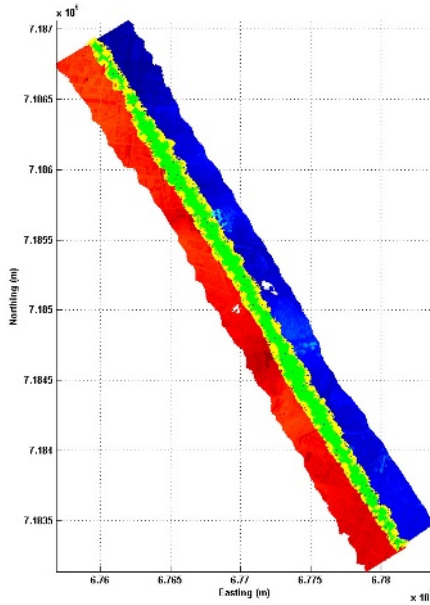


Fig. 5. Extracted ALS data points used for the adjustment process. Red and blue points represent overlapping ALS data strips, while yellow points lie within the overlapping area and green points are those used for the adjustment process.

and the process was repeated. Figure 6 shows the number of leaves of the CT as a function of the translation value in the X-direction. The value of translation in the X-direction is 0.41 when the number of leaves of the CT is minimized. Table 1 lists the parameters of transformation that were determined by the proposed algorithm. After applying the determined transformation, the total number of leaves of the CT in the overlapping area decreased from 150,178 to 141,112. Figure 7 shows the topography of the terrain before and after transformation; the number of leaves decreased following transformation. In the figure, the change in the number of leaves is only slight because this area is very small compared with the area as a whole. The number of leaves shown in Figure 7 decreased from 641 to 478, while the total number of points in this area is 3,599.

Table 1. Estimated parameters for 3-D conformal transformation

Parameter	Adjusted Value
S (scale)	1.00
X_T (m)	0.41
Y_T (m)	0.19
Z_T (m)	0.00
ω (degrees)	0.0e-7
ϕ (degrees)	-3.0e-7
κ (degrees)	-1.0e-7

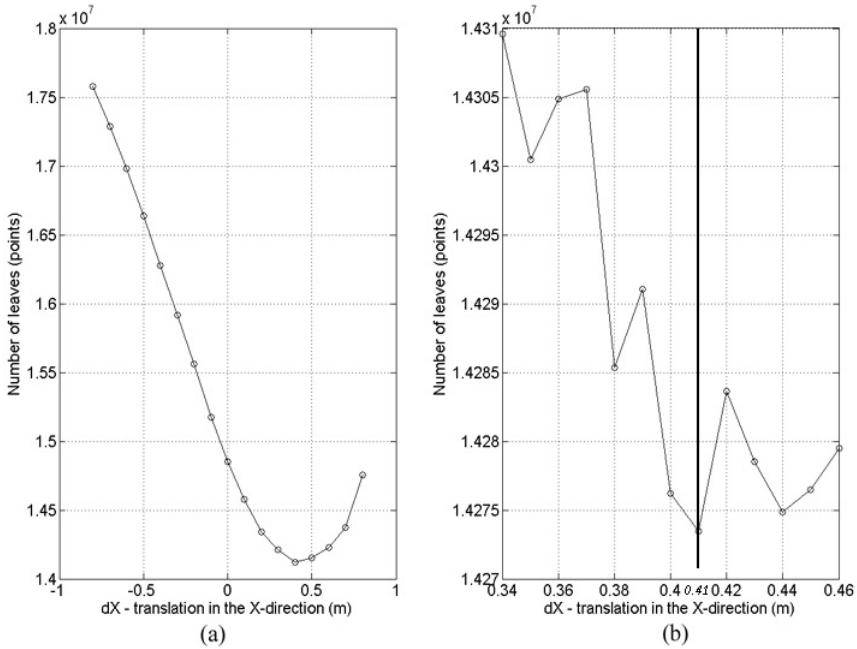


Fig. 6. Number of leaves of the CT as a function of the translation parameter in the X-direction. (a) First iteration to determine the translation value in the X-direction. (b) Second iteration to determine the translation value in the X-direction. When $x = 0.41$, the number of leaves is minimized (red data point)

To test the feasibility of the determined 3-D conformal transformation, a total of 164 pairs of conjugate linear features were identified and extracted from the two overlapping strips. These linear features were used to measure discrepancies before and after applying the transformation. In Figure 8, blue and red lines show conjugate linear features extracted from the two strips (see [7] for a complete description of the extraction process). The normal vector between each pair of linear features was used to measure discrepancies. For each pair of conjugate features, a normal vector was calculated from the midpoint of one conjugate linear feature to the other line. Table 2 lists the overall discrepancies between the two strips before and after applying the transformation. The means and standard deviations of the normal vectors indicate the existence of discrepancies between the strips before applying the transformation, especially in the X- and Y-directions. Following the transformation, it is evident that the discrepancies had been reduced, especially in the X- and Y-directions.

We next conducted a hypothesis test to determine whether the results are statistically significant. Using the paired comparison method, the differences between normal vectors before and after the transformation were examined to determine if they are significantly large from a statistical viewpoint. The test statistic was set up as follows:

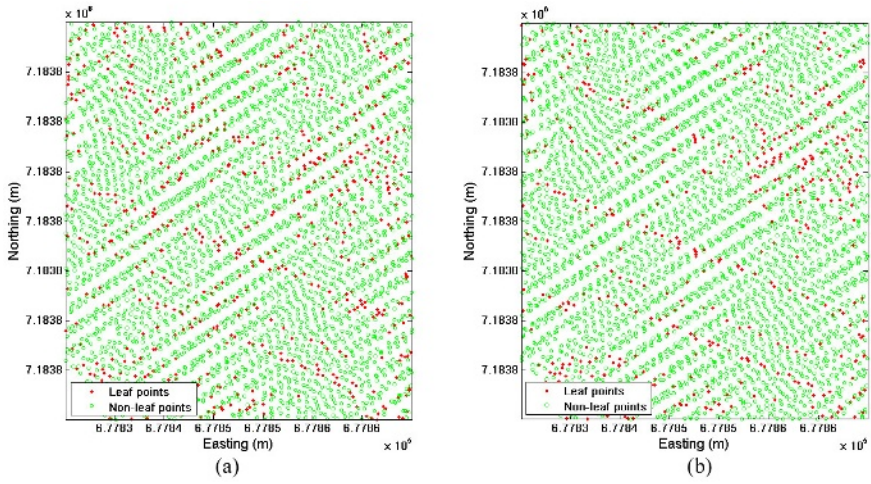


Fig. 7. Plot of a sub-area of the overlapping strips showing the change in the number of leaves of the CT before and after transformation. (a) Before transformation. (b) After transformation.

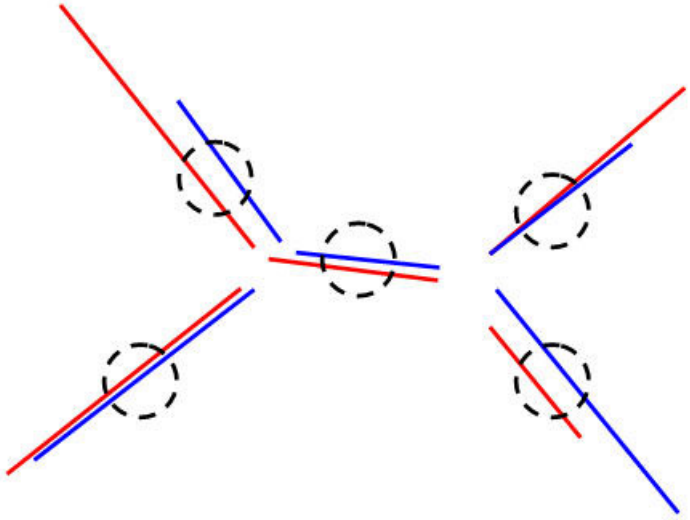


Fig. 8. Red and blue lines show five pairs of conjugate linear features extracted from two different strips

$$T = \frac{\frac{1}{n} \sum_{i=1}^n (X_{1,i} - X_{2,i}) - \delta_0}{S_D / \sqrt{n}} \tag{2}$$

where $S_D^2 = \frac{1}{n-1} \sum_{i=1}^n [(X_{1,i} - X_{2,i}) - \frac{1}{n} \sum_{i=1}^n (X_{1,i} - X_{2,i})]^2$ indicates the pooled standard deviation, $X_{1,i}$ and $X_{2,i}$ denote the value of i^{th} the normal vectors

before and after transformation, respectively, and n indicates the number of normal vectors.

In this case, σ_0 is zero. The corresponding hypothesis is:

$$H_0 : \mu_1 = \mu_2, H_1 : \mu_1 > \mu_2 \quad (3)$$

where μ_1 and μ_2 are the population means of the normal vectors before and after applying the transformation, respectively.

The test results reject the null hypothesis at the 99% significance level. The t-values are 7.245 in the X-direction and 7.143 in the Y-direction, respectively, which rejects the null hypothesis at the significance level (99%, at which the t-value is 2.364). This result indicates that the transformation reduced the discrepancies between the ALS data strips by a statistically significant amount in both the X- and Y-directions.

Table 2. Measurements of discrepancies before and after applying the transformation

	Before Transformation		After Transformation	
	Mean	Standard Deviation	Mean	Standard Deviation
dx(m)	-0.182	±0.286	-0.028	±0.219
dy(m)	-0.082	±0.315	-0.009	±0.166
dz(m)	0.003	±0.111	-0.015	±0.070

5 Conclusions and Future Work

This paper presents a generally applicable algorithm that adjusts for discrepancies between ALS data strips. The method overcomes the limitations involved in methods that use conjugate features in the adjustment process. By using a CT algorithm and the topological characteristics of the terrain, the algorithm explicitly formulates step-by-step methodologies to determine the most suitable transformation for adjustment. We applied the method to an ALS dataset and achieved a statistically significant reduction in discrepancies, especially in the X- and Y-directions. We are now focusing on the way in which the results are affected by the point density of ALS data and the number of points used for adjustment. We are also investigating a method of extracting sub-areas from overlapping area based on the distribution and terrain type of a sub-area; this will improve computation efficiency. In addition, for general applications, this algorithm should be tested for specific types of terrain such as very flat terrain and vegetated terrain.

Acknowledgments. The authors thank to financial support from the Korea Aerospace Research Institute and the Engineering Research Institute, Seoul National University.

References

1. Pfeifer, N.: Airborne Laser Scanning Strip Adjustment and Automation of Tie Surface Measurement. *Boletim de Ciências Geodésicas*. Vol. **11** (2005)
2. Kager, H., Kraus, K.: Height Discrepancies between Overlapping Laser Scanner Strips. *Proceedings of Optical 3D Measurement Techniques V*. Vienna. Austria (2001) 103-110
3. Maas, H.G.: Least-Squares Matching with Airborne Laser Scanning Data in a TIN Structure. *International Archives of Photogrammetry and Remote Sensing* **33(B3/1)** (2000) 548-555
4. Filin, S., Vosselman, G.: Adjustment of Airborne Laser Altimetry Strips. *International Archives of Photogrammetry and Remote Sensing*, Vol. XXXV, B3, Istanbul, Turkey (2004)
5. Kager, H.: Discrepancies between Overlapping Laser Scanner Strips-Simultaneous Fitting of Aerial Laser Scanner Strips. *XXth ISPRS Congress Vol. XXXV part B/1*. Istanbul. Turkey (2004)
6. Vosselman, G.: On the Estimation of Planimetric Offsets in Laser Altimetry Data. *International Archives of Photogrammetry and Remote Sensing*. Vol. XXXIV 3A. Graz. Austria (2002) 375-380
7. Lee, J., Yu, K., Kim, Y., Habib, A.F.: Segmentation and Extraction of Linear Features for Adjustment of Discrepancies between ALS Data Strips. *IEEE Proceedings IGARSS 2005*. Seoul. Korea (2005)
8. Pascucci, V., Cole-McLaughlin, K.: Efficient Computation of the Topology of Level Sets. *IEEE Proceedings Visualization 2002* (2002)
9. Bollobas B.: *Modern Graph Theory*. Springer-Verlag. New York (1998)
10. Shinagawa, Y., Kunii, T.L., Kergosien, Y.L.: Surface Coding Based on Morse Theory. *IEEE Computer Graphics and Applications* **11** (1991) 66-78
11. Boyell, R.L., Ruston, H.: Hybrid Techniques for Real-time Radar Simulation. *IEEE Proceedings Fall Joint Computer Conference 63*. Las Vegas. USA (1963) 445-458
12. Bajaj, C.L., Pascucci, V., Schikore, D.R.: The Contour Spectrum. *IEEE Proceedings Visualization 1997* (1997) 167-175
13. Carr, H., Snoeyink, J., Axen, U.: Computing Contour Trees in All Dimensions. *Computational Geometry* **24(2)** (2003) 75-94

Visual Bootstrapping for Unsupervised Symbol Grounding

Josef Kittler, Mikhail Shevchenko, and David Windridge

Center for Vision, Speech and Signal Processing,
University of Surrey, Guildford,
GU2 7XH, United Kingdom

{j.kittler, m.shevchenko, d.windridge}@surrey.ac.uk

Abstract. Most existing cognitive architectures integrate computer vision and symbolic reasoning. However, there is still a gap between low-level scene representations (signals) and abstract symbols. Manually attaching, i.e. grounding, the symbols on the physical context makes it impossible to expand system capabilities by learning new concepts. This paper presents a visual bootstrapping approach for the unsupervised symbol grounding. The method is based on a recursive clustering of a perceptual category domain controlled by goal acquisition from the visual environment. The novelty of the method consists in division of goals into the classes of parameter goal, invariant goal and context goal. The proposed system exhibits incremental learning in such a manner as to allow effective transferable representation of high-level concepts.

1 Introduction

The field of Artificial Cognitive Systems within Artificial Intelligence developed over the last decade with the intention of constructing intelligent systems with abilities to perceive, learn, communicate and understand the external world in the manner of biological organisms. This approach draws on and integrates methods of computer vision, neural networks and symbolic AI (Granlund 2005, Sun and Wermter 2000). Perceptual information, usually obtained by machine vision, is processed for learning low-level models using, for example, neural network algorithms (Sommer 2005). These models are then transferred to the symbolic level for further abstract reasoning by means of classical AI techniques such as Computer Linguistics.

The diversity of existing research is manifested by various methods of representing the perceived data, training the networks and generalizing symbols at the top hierarchical level of the system. Hence, there is significant progress in developing both connectionist and symbolic approaches; however, the key issue of linking them together has not yet been solved. The problem thus arises of how low-level models are to be transferred into the abstract representation (and, equally, how symbols are to be referred down to the neural network states).

In the field of cognitivism this problem is called symbol grounding (sometimes symbol attachment) and deals with giving the symbols their physical meaning

(Harnad 2002, Sloman 2005). The existing models of cognitive systems perform symbol grounding manually, by user interruption, causing the known problems of pre-hardwired knowledge, non-expanding competences and absence of autonomous exploration. The goal of our research is to provide the cognitive architecture with the mechanism of unsupervised symbol grounding (USSG).

The main approach for USSG is based on automatic clustering of incoming low-level information. For reasonably complex perception any automatic clustering is not reliable since it fails to distinguish among categories (Harnad 2003). We propose a visual bootstrapping approach to symbol grounding when the system starts from simple low-level models building up primitive symbolic categories and reusing them on further level of complexity (Granlund 2005). Recursive clustering of previously obtained categories on each next level of abstraction brings out incremental developing of hierarchical symbolic structure. The algorithm of learning behavioural models of detected symbols is based on the Markov Decision Process and its extension to the Reinforcement Learning (Gullapalli 1992). The difference from the standard methods is that the system reward is not a predefined set of scalars but a parameter calculated online using a visual distance to the goal. The innovation of our approach is that we introduce three different classes of goal representation: elementary, invariant and context. An *elementary goal* takes only low-level visual features of the significant percepts. It can be transferred into a *invariant goal* used at the abstract level. A *context goal* is a projection of the invariant goal onto the current visual context.

In section 2 we describe what the significant percepts are, a role of the goal in the bootstrapping and the process of goal detection in our method. Section 3 is devoted to the problem of unsupervised path-finding in order to attain the goal by action. It proposes mechanisms of learning essential invariant perception-action couples and establishing models of system active behaviour. Section 4 presents a learning scenario where the method is applied for the particular kind of visual environment. Finally, section 5 discusses the results and future work.

2 Goal Detection

Goal-oriented behaviour allows living organisms to address the high complexity of the perceptual information coming from the real world. The internal world representation should consist of only significant structures and events of the system's surrounding. That fact underlies the mechanism of autonomous goal acquisition. We define the system goal as a significant perceptual state detected in the visual environment.

Suppose that the system has a manipulator and a camera. The state of the manipulator can be described by a set of parameters $\{m_k\}$ and any action as a vector $\Delta\mathbf{M} = \{\delta m_k\}$. Also, the visual system transforms an input image into a vector of features $\mathbf{S} = \{f_i\}$ (In section 4 we will describe an algorithm for calculating the feature vector).

Since we imply no prior goal when the system starts, it performs an arbitrary behaviour by generating a set of random motor changes $\{\Delta\mathbf{M}_j\}$. The set of

states $\{\mathbf{S}_j\} = \{\{f_i\}_j\} = \{f_{ij}\}$ corresponding to the random actions is perceived by the visual system. The low-level visual features are integrated upon the frames in order to obtain frequency histograms $I(f_s)$ for each value of f_{ij} :

$$I(f_s) = \sum_{i,j} \begin{cases} 1 & f_s = f_{ij} \\ 0 & f_s \neq f_{ij} \end{cases} \quad i = 1 \dots n^{(j)}, \quad j = 1 \dots N. \quad (1)$$

The values with relatively high frequencies are considered as the elementary (parameter) goals since the peaks of the histogram do not represent the whole scene but only one low-level feature. A visual state with the given parameter goal with other feature vector components taken from the current visual scene is the context goal. A visual state with the given parameter goal and any other feature vector components create the invariant goal which does not depend on the current context. The invariant goal is used for representing the detected visual structure or event and the context goal implements its invariant components when the system applies the model to generate action. Several invariant goals can be combined into a *complex invariant goal* with the several parameter goals. It makes possible to take previously obtained models for generating new kinds of symbolic classes and correspondent novel behaviour.

3 Bootstrapping by Attaining Visual Goals

3.1 Problem Definition

Suppose that the system has detected an invariant goal and it is not equal to the current visual state \mathbf{S}_i^c . Any physical action $\Delta\mathbf{M}_i$ changes the state of the external world that, in its turn, modifies the current perceptual state \mathbf{S}_{i+1}^c . We denote this transformation which carries a sense of the physical model of the external world as L :

$$L: \{\mathcal{M}\} \rightarrow \{\mathcal{S}\}, \quad \mathbf{S}_{i+1}^c = L(\Delta\mathbf{M}_i). \quad (2)$$

A system internal transformation containing the learning mechanism is represented as another function mapping a perceptual state onto a system response (Figure 1):

$$Q: \{\mathcal{S}\} \rightarrow \{\mathcal{M}\}, \quad \Delta\mathbf{M}_{i+1} = Q(\mathbf{S}_{i+1}^c). \quad (3)$$

The objective of the algorithm is to find a function Q^g which generates an appropriate action moving the system towards the invariant goal state \mathbf{S}^g . Formally we can consider the task as following:

1. Prove that a policy $Q^g(\mathbf{S}^c)$ exists such that the sequence $\mathbf{S}_{[i]}^c$ converges on the goal $\mathbf{S}_{[i]}^c \rightarrow \mathbf{S}^g$ for a finite number of steps i with given L
2. Find a policy $Q^g(\mathbf{S}^c)$ such that the sequence $\mathbf{S}_{[i]}^c$ converges on the goal $\mathbf{S}_{[i]}^c \rightarrow \mathbf{S}^g$ for a minimal number of steps i , any initial state $\mathbf{S}_{[0]}^c$ and given L

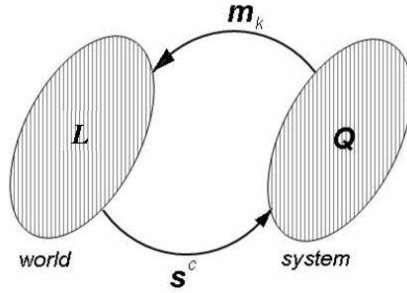


Fig. 1. Transformations L and Q

The default model for Q^g is based on repeating successful actions obtained by random trials. The first movement is purely random:

$$Q^g(S_0^c) \in \{\mathcal{M}\}, \quad \forall S_0^c \in \{\mathcal{S}\}. \tag{4}$$

If the trial is followed by decreasing of the invariant goal distance then, on the next step, the system perform the same action; if not, then $Q^g(S_1^c)$ will be chosen randomly again. For any i :

$$Q^g(S_{i+1}^c) = \begin{cases} \in \{\mathcal{M}\}, & D_i \geq D_{i-1} \\ Q^g(S_i^c), & D_i < D_{i-1} \end{cases} \tag{5}$$

$$D_i = D(S_i^c, S^g). \tag{6}$$

The process goes on until the distance D becomes small

$$D_K = D(S_K^c, S^g) \leq \varepsilon. \tag{7}$$

The sequence of the movements and the corresponding perceptual states converging on the goal provide the system with a set of samples that form a primitive mapping from perception to action. Our objective is to find a function which transforms any percept into the movement. The transformation must be optimal and invariant to the starting visual state.

The process of finding such a policy has two stages. Firstly, the most significant visual states are detected by calculating their frequencies (or probabilities) within the sample sequence. The selected key states build up an initial policy Q^g . Secondly, Q^g is being refined by setting up other experiments, obtaining new samples and updating the state probabilities in order to find the states which do not depend on the starting configuration.

3.2 Acquiring a Primitive Model

Suppose that P_r is a probability to find a perceptual state S_r^c in a sequence $S_{[i]}^c, i = 1 \dots N$, such that

$$P_r = \sum_{i=1}^K (S_r^c = S_i^c) / N. \tag{8}$$

The visual states having high values of P_r are treated as the significant ones and added to the model keystate list $\{\mathbf{S}_t^0\}$:

$$P_r > \delta : \quad \mathbf{S}_t^0 = \mathbf{S}_r^c, \quad \mathbf{S}_r^c \in \{\mathbf{S}_i^c\} \tag{9}$$

where δ is the threshold of "significance".

The response of policy Q^g is defined by the sample $Q^g(\mathbf{S}_a^c)$ if there exists an index a which satisfies the following:

$$\begin{cases} D(\mathbf{S}_a^0, \mathbf{S}^g) < D(\mathbf{S}^c, \mathbf{S}^g) \\ \min_{a=1\dots T} \{D(\mathbf{S}^c, \mathbf{S}_a^0)\} \end{cases} \tag{10}$$

If such an index is not found then the response is the action corresponding to the goal state:

$$Q^g(\mathbf{S}^c) = \begin{cases} Q^g(\mathbf{S}_a^0), & \exists a \\ Q^g(\mathbf{S}^g), & \nexists a \end{cases} \tag{11}$$

Obviously, the state probabilities for the first run give the same value for any visual state in the sample sequence. All of them are taken as the keystates defining the initial model of perception-action transformation. This can already be considered an improvement since implementation of Q^g converges on the goal. During the following series of runs with different starting configurations the policy will be refined by updating the model keystate list.

3.3 Improving the Model

Let's suppose that after the first trial the primitive model Q_0^g has been sampled¹. We also have the current perceptual state \mathbf{S}_0^c which is the starting point of the next experiment. Since the system already has a model, even if it is a primitive one, the first movement is not taken randomly – it is the corresponding value of $Q_0^g(\mathbf{S}_0^c)$:

$$Q_1^g(\mathbf{S}_0^c) = Q_0^g(\mathbf{S}_0^c). \tag{12}$$

The rest of the procedure which takes the sample $Q_1^g(\mathbf{S}_i^c)$ is the same as for the first experiment (see eq. 10,11) except that the previously obtained model is used instead of random trials:

$$Q_1^g(\mathbf{S}_{i+1}^c) = \begin{cases} Q_0^g(\mathbf{S}_i^c), & D_i \geq D_{i-1} \\ Q_1^g(\mathbf{S}_i^c), & D_i < D_{i-1} \end{cases} \tag{13}$$

The model update is done by recalculating the state probabilities for the current sample within distribution P_r :

$$P_r = \frac{P_{r,0} + P_{r,1}}{2} \tag{14}$$

where $P_{r,0}$ and $P_{r,1}$ are the state probability distributions for the first and second experiments respectively.

¹ New indexing for policy Q^g is introduced her to denote the current run.

Without losing generality we can write the algorithm of model update for the n^{th} run:

$$Q_n^g(\mathbf{S}_0^c) = Q_{n-1}^g(\mathbf{S}_0^c) \quad (15)$$

$$Q_n^g(\mathbf{S}_{i+1}^c) = \begin{cases} Q_{n-1}^g(\mathbf{S}_i^c), & D_i \geq D_{i-1} \\ Q_n^g(\mathbf{S}_i^c), & D_i < D_{i-1} \end{cases} \quad (16)$$

$$P_r = \frac{P_{r,n-1} + P_{r,n}}{2}. \quad (17)$$

4 Learning Scenarios and Experiments

4.1 The Perception-Action Level

We carried out our experiments on software simulating the physical "world" as well as the system itself. The world is a 2D square box which has boundaries that restrict movements of the manipulator. The system motor domain has four DOFs defining position of the arm: the length R , the angle ϕ , the gripper angle θ , the gripper state γ . The visual system performs attractor detection, visual attractor description, organizing the visual memory and recognition. The mechanism of discovering attractors is based on motion detection and tracking. It eliminates static, unknown "background" objects from the processing. For instance, during the random exploration mode this mechanism takes into account only the manipulator if it is the only object moving on the scene. It also can be other objects moved by the robot arm or a user. The visual scene is represented by a graph, each vertex of which is an attractor feature vector with the following components: the attractor id, the positions in Cartesian coordinates x and y , the attractor orientation α and the changes of the position and orientation after the last action $dx, dy, d\alpha$ (figure 2):

$$\mathbf{S}^c = \{\mathbf{f}_j\} = \{(id, x, y, \alpha, dx, dy, d\alpha)_j\}. \quad (18)$$

The visual distance between two scenes is a normalized sum of distances among the attractors

$$D(\mathbf{S}_i^c, \mathbf{S}_n^c) = \sum_{i,k=1}^{L,N} d(i, k) / (L \cdot N) \quad (19)$$

where L, N are the numbers of the attractors for the scenes \mathbf{S}_i^c and \mathbf{S}_n^c respectively.

The attractor distance is calculated as a weighted sum of the distances between corresponding components of the feature vectors:

$$d(i, k) = \delta id + \delta C + \delta A \quad (20)$$

$$\delta id = \begin{cases} 0 & id_i = id_k \\ 1 & id_i \neq id_k \end{cases} \quad (21)$$

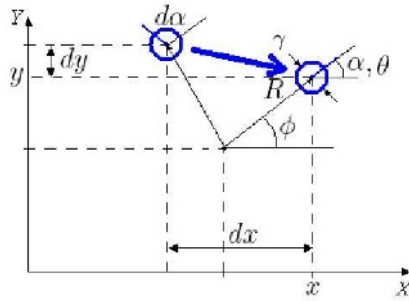


Fig. 2. Motor and visual parameters

$$\delta C = \frac{\sqrt{(x_i - x_k)^2 + (y_i - y_k)^2 + (dx_i - dx_k)^2 + (dy_i - dy_k)^2}}{\sqrt{2X^2 + 2Y^2}} \tag{22}$$

$$\delta A = \frac{\sqrt{(\alpha_i - \alpha_k)^2 + (d\alpha_i - d\alpha_k)^2}}{2\sqrt{2}\pi} \tag{23}$$

where

$$\mathbf{f}_i = (id_i, x_i, y_i, \alpha_i, dx_i, dy_i, d\alpha_i) \tag{24}$$

$$\mathbf{f}_k = (id_k, x_k, y_k, \alpha_k, dx_k, dy_k, d\alpha_k), \tag{25}$$

and where X, Y are the horizontal and vertical sizes of the workspace.

The elementary goal is calculated within a short-term visual memory. It stores the scene descriptors for up to 30 frames; each frame is taken after a movement has been detected on the scene. The detected goals are converted into the invariant representation, stored in a permanent memory and linked with the activity models Q^g obtained after the consequent learning.

4.2 Obtaining Local Motor Control

Suppose that the system starts movement in the random mode. Only the manipulator is detected as a visual object on the scene. The random generator produces small changes of the motor parameters and we consider the resulting movements as local ones. The precision of measuring the corresponding local changes on the visual scene is low; therefore we are allowed to quantize visual movements. Let us define four directions within a local surrounding of the manipulator position (see figure 3). Any local movement is perceived as one of the four quantum steps of the closest direction and the unity length. It is obvious that after a series of small random movements all the quantum steps will be detected as the significant events (since the corresponding values of dx or dy are constant) and each of those movements becomes the goal. Let us consider the quantum movement $q_r, r = 1, \dots, 4$ as a current parameter goal. According to the method of finding

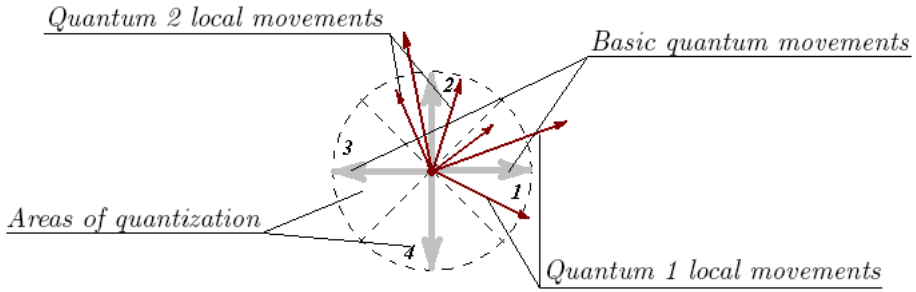


Fig. 3. Visual quantization of local movements

the motor solution for the invariant goal f^g derived from q_r the system obtains a set of perception-action couples $\{f_j, \delta m_i\}_k$ and the transformation Q^g defining the motor response for the given visual input (see eq. 15-16). The example of learning motor control of moving the manipulator right (the quantum step 1) is shown in figure 4. The samples are from various starting configurations and the trajectories demonstrate how well the model describes arm control on each stage, namely, at the beginning, after 10 movements to the right boundary, after 20 movements to the right boundary. On the global scale even the best trajectories do not strictly follow the horizontal line sometimes because Q_g does not return an appropriate response. The system switches to the random exploration mode to find the needed motor changes and update the current model. But most of the time local movements belong to the chosen quantum and the trajectories are explained by the errors of the local visual measurements.

Using the simulator the same method has been applied to learn control of local gripper rotation and grasping movements.

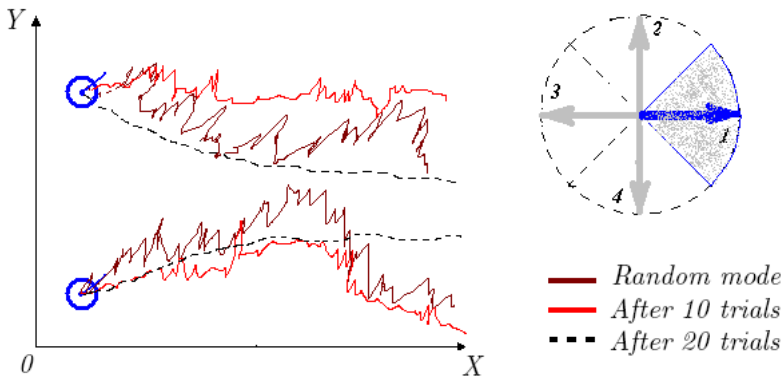


Fig. 4. Local horizontal movements

4.3 Learning Global Motor Control

Let us add an object to the scene. Visually the system detects another attractor and the spatial relation between the arm and the object is:

$$S^c = \{f_1, f_2, f_{1,2}\} \quad (26)$$

$$f_{1,2} = d(1, 2). \quad (27)$$

If the system frequently finds that its manipulator is on the same position as the object then a new parameter goal is detected. The corresponding behavioural task is the intention to attain the object position by the manipulator. On this level of representation, instead of the parameters directly controlling different motors in the motor domain, the system operates within the set of previously acquired competences - the local movement models: $\{\delta m_k\} \equiv Q_k^g$. Applying the learning mechanism (eq. 15-17) the system obtains the model of approaching the object from any starting position. The example of the system performance in the random mode, after 10 and 20 training series, is demonstrated in figure 5. Other visual goals detected on this level, and the corresponding learned models of the system response, are grasping the object and moving the object.

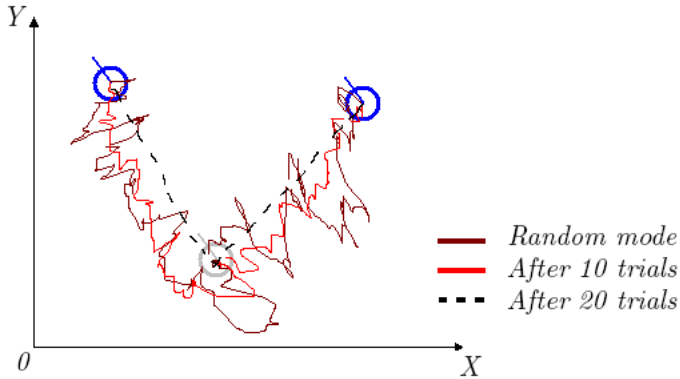


Fig. 5. Approaching an object from different starting positions

5 Conclusions

We have presented a method of detecting perceptual categories in an unsupervised manner by visual bootstrapping. The learned categories have features that are invariant to the scene context, letting us consider them as symbolic tokens. The principles of autonomous goal acquisition, random trials and iterative repeating of successful actions control the process of detecting symbols. Using previously obtained models instead of the highly complex low-level perceptual information makes categorization reliable. What the categories detected

by the system is defined by the significant events in the visual environment, not pre-hardwired by the user. The experimental section demonstrates how such a system can be bootstrapped up to the level of manipulator control. Further research will be carried out in the direction of learning ways of arranging objects according to particular rules. One possible scenario is a shape-sorter game, where the task is to insert various blocks into holes of corresponding shapes. The system operating in such an environment detects visual goals and builds up the new levels of symbolic hierarchy implementing the game rules. This demonstration will prove the system's ability not only to create primitive competences such as arm or object movement control but also to understand complex world events and generate high-level behaviour.

References

1. Billard, A. and Dautenhahn, K.: Experiments in social robotics: grounding and use of communication in autonomous agents, *Adaptive Behavior*, **7(3/4)**, (2000), 415–438.
2. Granlund, G.: A Cognitive Vision Architecture Integrating Neural Networks with Symbolic Processing, *Kunstliche Intelligenz*, **2**, (2005), 18–24.
3. Harnad, S.: Symbol Grounding and the Origin of Language. In *Computationalism: New Directions*, MIT Press, (2002), 143–158.
4. Harnad, S.: Categorical Perception. *Encyclopedia of Cognitive Science*. Nature Publishing Group, Macmillan, (2003).
5. Li, H. Olver, J. and Sommer, G.: Computer Algebra and Geometric Algebra with Applications. In *Proceedings IWMM 2004 and GIAE 2004*, Springer-Verlag Berlin Heidelberg, LNCS 3519, (2005), 258–277.
6. Fodor, J. and Pylyshyn, Z.: Connectionism and cognitive architecture: A critique. *Cognition*, **28**, 1988, 3–71.
7. Gullapalli, V.: Reinforcement learning and its application to control. Ph.D. thesis, University of Massachusetts, Amherst, MA, (1992).
8. Pauli, J. and Sommer, G.: Perceptual organization with image formation compatibilities. *Pattern Recognition Letters*, **23**, 2002, 803–817.
9. Sloman, A. and Chappel, J.: The altricial-precocial spectrum for robots. In *proceedings IJCAI'05*, (2005).
10. Wermter, S and Sun, R.: *Hybrid Neural Systems*. Springer, Heidelberg, New York, (2000).

A 3D Model Acquisition System Based on a Sequence of Projected Level Curves

Huei-Yung Lin, Ming-Liang Wang, and Ping-Hsiu Yu

Department of Electrical Engineering, National Chung Cheng University, 168 University Rd., Min-Hsiung, Chia-Yi 621, Taiwan, R.O.C.

lin@ee.ccu.edu.tw, pool11z.wang@msa.hinet.net, u9042036@ccu.edu.tw

Abstract. An image-based 3D model acquisition system using projections of level curves is presented. The basic idea is similar to surface from parallel planar contours, such as 3D reconstruction from CT or laser range scanning techniques. However, our approach is implemented on a low-cost passive camera system. The object is placed in a water container and the level curves of the object's surface are generated by raising the water level. The 3D surface is recovered by multiple 2D projections of parallel level curves and the camera parameters. Experimental results are presented for both computer simulated data and real image sequences.

1 Introduction

3D model acquisition of real world objects is an active research topic in the areas of computer vision, CAD/CAM, and pattern recognition. The applications of 3D computer models range from reverse engineering and industrial inspection to computer graphics and virtual reality. With a widely adopted passive camera system, commonly used 3D shape recovery techniques include stereo vision or structure from motion, shape from shading, shape from silhouettes, and photometric stereo, etc [1,2]. These approaches usually require either multiple images captured from different viewpoints or controlled illumination conditions for image acquisition. There are also some other techniques such as depth recovery from zooming/focus/defocus, which extract the depth information by comparing several images recorded by a single camera with different camera parameter settings [3,4]. A motorized zoom lens is required to change the zoom or focus positions for these methods, and elaborate camera calibration generally has to be carried out first.

In addition to the conventional image-based methods, 3D model reconstruction of real objects can also be achieved by means of active sensors. Laser range scanning, structured lighting, computed tomography (CT scan), and magnetic resonance imaging (MRI) are several popular approaches. These techniques usually process the object's surface one cross-section at a time to obtain the corresponding 3D curve, and then merge the layered information into a complete 3D surface. Generally speaking, 3D model acquisition using active sensors provides more accurate results than those derived from passive camera systems. However,

the expensive equipment, elaborate system calibration and setup have restricted their applications mostly in the laboratory environments.

Inspired by the shape recovery techniques based on multiple planar scans of the object, in this work we propose a 3D model acquisition system using the level curves associated with the object's cross-sections. The idea is to use a passive camera system to capture multiple 2D projections of parallel planar 3D curves, and recover the corresponding level curves of the object surface. The 3D surface of the object can then be obtained by combining the multiple 3D curves or contours. Our current research is focused on developing an image-based 3D model acquisition system since the cameras are ubiquitous and relatively inexpensive. In addition to the cost and performance issues, another important design principle is how to make the system easy to implement and use. More specifically, the calibration of the vision system should be carried out with least human interference.

In our prototype 3D model acquisition system, the test object is placed in a water container mounted on a computer controlled turntable. The image sequences are captured by a static camera located in front of the rotation stage. Different from most active 3D reconstruction techniques using the projection of known patterns on the object's surface, the level curves acquired in our system are generated by increasing the water level in the container. For a given viewpoint, the visible surface of the object is recovered using projections of the parallel level curves and the camera parameters of the vision system. Furthermore, the complete 3D model (full 360° view) can be reconstructed by data registration and integration of multi-view 3D shape acquisition from different viewpoints.

Although 3D model reconstruction from parallel slices of contours is a well developed technique [5,6], it usually requires expensive data acquisition equipment (such as CT or MRI systems). The data collection process also has to be performed under specific environments. As for the structured lighting or laser range scanning systems, the reflectance property of the visible surface is one important issue to be solved, especially the active lighting or projection near the edges or abrupt depth changes of the object surface. The proposed 3D model acquisition method uses low-cost hardware setup, and it is relatively insensitive to the environmental illumination change. Experimental results are presented for both computer simulated data and real image sequences.

2 Shape from Parallel Planar 3D Curves

3D surface recovery of the proposed method is based on the integration of a set of 3D curves, with each of them lies on one of a series of parallel planes. As shown in Fig. 1, suppose a cross-section of an object is given by the intersection with a plane

$$\mathbf{n}^T \mathbf{x} = d \tag{1}$$

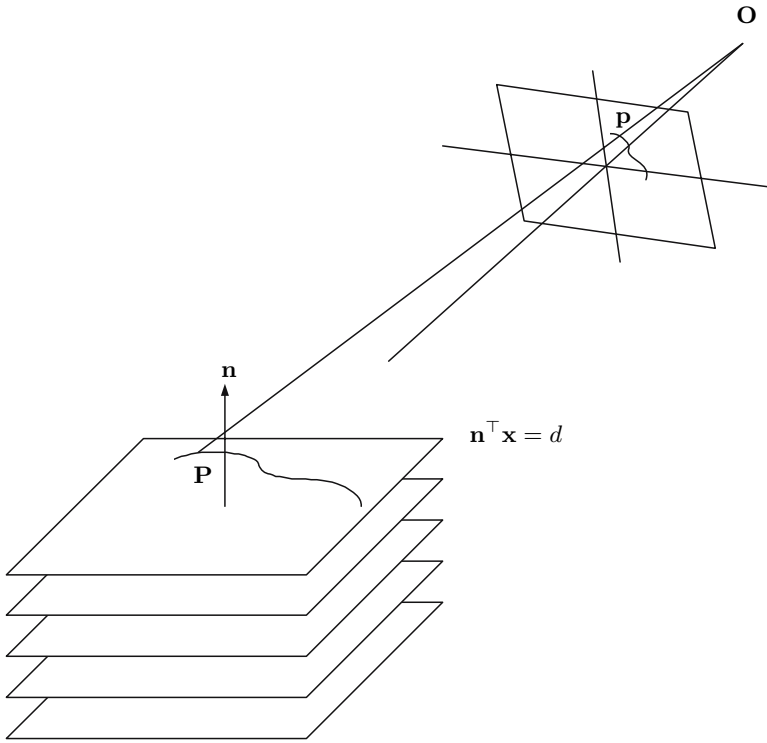


Fig. 1. Parallel planar 3D curves and the camera model

where \mathbf{n} is the plane normal. Let the normal vector \mathbf{n} be represented by (a, b, c) , then the relationship between a point (x, y, z) on the 3D curve from the cross-section and its image point (\hat{x}, \hat{y}) is given by the perspective projection

$$x = \lambda \hat{x}, \quad y = \lambda \hat{y}, \quad z = \lambda f \tag{2}$$

with the scale factor

$$\lambda = \frac{d}{a\hat{x} + b\hat{y} + cf} \tag{3}$$

where f is the focal length of the camera. Thus, it is clear that the coplanar 3D curve can be recovered from its projection on the image plane for a given set of parameters (a, b, c, d) by Eqs. (2) and (3).

Based on the above derivation, the viewable surface of an object from a single viewpoint can be reconstructed by stacking the 3D curves of the cross-sections obtained from the intersections with parallel planes. If the recovered 3D curves are *simple*, i.e. there are no self-intersections and self-occlusions, each level curve on cross-section i can be represented by a parameterized curve $\mathbf{c}_i(t)$, where $t \in [0, 1]$. For a given set of parallel planar 3D curves, the surface mesh can then be generated by connecting the vertices (i.e. the recovered 3D points from the

image pixels) on the parameterized curves of two consecutive cross-sections $\mathbf{c}_i(t)$ and $\mathbf{c}_{i+1}(t)$. More sophisticated surfaces from contours approaches [7,8] can be adopted for more general cases. However, it should be noted that there are some fundamental limitations on range data acquisition from single viewpoint, which makes the simple curve assumption readily available for the imaging system.

It is not possible to calculate the parameters a, b, c, d of the plane equation (1) from the corresponding image without any metric information. Thus, a camera pose estimation method based on the projection of a rectangle (with known size) on the plane is adopted [9]. In the implementation, this rectangular shape can be easily obtained from the planar surface patch bounded by the cubic water container. The detailed derivation is given as follows.

It is shown that the relative depths of four 3D points can be determined by their 2D projections if they form a parallelogram in the 3D space [10]. Thus, given the image points of the rectangular surface patch, the corresponding 3D points can be computed up to a unknown scale factor. Now, suppose the four corner points in the 3D space are $P_i = (x_i, y_i, z_i)$, and the corresponding image points are $p_i = (\hat{x}_i, \hat{y}_i, f)$, where $i = 0, 1, 2, 3$, and f is the focal length of the camera. Then we have $(x_i, y_i, z_i) = (\lambda_i x_i, \lambda_i y_i, \lambda_i f)$, where λ_i 's are the unknown scale factors. Since the relative depths of the corner points can be written as $\mu_i = \lambda_i/\lambda_0$, for $i = 1, 2, 3$, we have

$$\begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} = \begin{pmatrix} x_1 - x_2 & x_3 \\ y_1 - y_2 & y_3 \\ 1 & -1 & 1 \end{pmatrix}^{-1} \begin{pmatrix} x_0 \\ y_0 \\ 1 \end{pmatrix} \tag{4}$$

If the dimension of the rectangular surface patch is known, say, the width is W , then $W^2 = |P_0 - P_1|^2$. Consequently, we have

$$\lambda_0 = \frac{W}{\sqrt{(x_0 - \mu_1 x_1)^2 + (y_0 - \mu_1 y_1)^2 + f^2(1 - \mu_1)^2}} \tag{5}$$

That is, λ_0 can be calculated using μ_1 given by Eq. (4), and then λ_i 's can be obtained by the relative depths, $\mu_i = \lambda_i/\lambda_0$, for $i = 1, 2, 3$. Finally, the parameters a, b, c, d used in Eq. (3) can be determined by the image of the corner points associated with a given water level and the dimension of the cubic container.

3 Simulation with Synthetic Data Sets

To verify the correctness of the proposed 3D model reconstruction method from parallel planar contours, computer simulation with synthetic data set is carried out first. Several computer generated 3D models are used as test objects, and the coplanar level curves are given by the intersections of a 3D computer model and a sequence of predefined parallel planes. A virtual camera is used to capture the projections of the visible level curves of the 3D model surface associated with

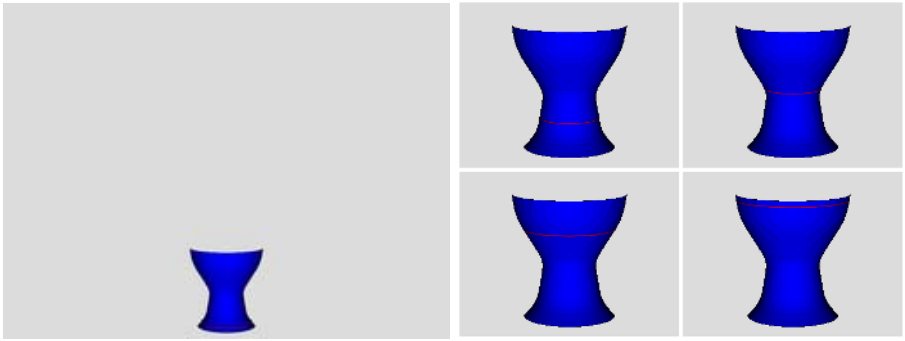
the given viewpoint. The computer generated 3D model is then rotated with respect to an axis inside the object for every 45° to capture the images of level curves from different viewing directions. In the simulation the parallel planes are created with equally spaced distance, however, it is not a general requirement for the proposed method.

The simulation results of two computer generated objects, a vase and a cube, are illustrated in Fig. 2. The Visualization Toolkit (VTK) is used for both image acquisition of the ideal 3D models and rendering of the reconstructed 3D models [11]. Figs. 2(a) and 2(b) show the images of visible surface with level curves captured by the virtual camera. In the simulation the optical axis of the camera is parallel to the cross-sections of the object. Thus, the camera is placed much higher than the object to ensure perspective of the level curves. The originally captured images are shown in the left figures, and the right figures show the close-up with more detailed information. There are 20 and 30 slices of level curves generated for the vase and cube object, respectively. Only a few of them are shown in the figures. The 3D curves are recovered sequentially from the top, using their corresponding edge segments detected in the image. Fig. 2(c) shows the reconstruction results from different viewpoints. 20 and 30 slices of cross-sections are demonstrated for these two cases, and better 3D recovery can be achieved by increasing the number of level curves.

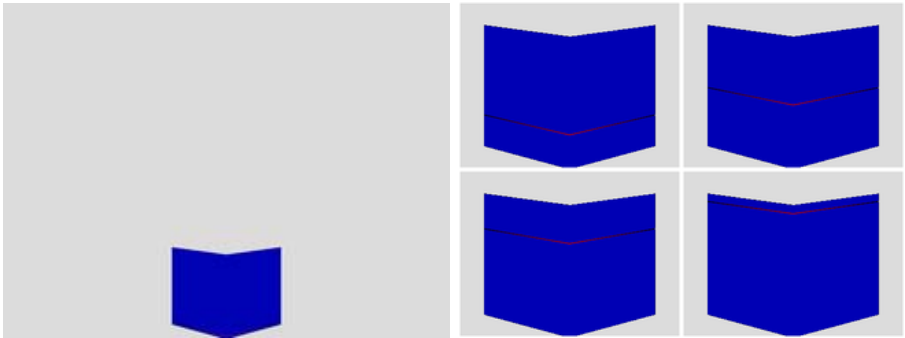
4 Vision System Design and Implementation

For the 3D model acquisition of real objects, the schematic diagram of the proposed prototype system is illustrated in Fig. 3(a). The cubic container used to place the object is specifically designed so that the water level can increase smoothly from the equally spaced holes on the bottom. As depicted in Fig. 3(b), the second and first layers are used to create a buffer zone for smooth water input and output transition. The container is placed on a PC controlled turntable for multi-view 3D shape recovery with a single camera. For a given viewpoint, the problem of determining the plane equation of the water level for 3D reconstruction is equivalent to the problem of camera pose estimation. A self-calibration method using the water level curves is described in Section 2. The focal length is given by the camera setting. As for the complete 3D model acquisition, it is also mandatory to find the rotation axis of the turntable for multi-view 3D data registration and integration. To avoid additional calibration for the rotation axis, the unit vector and location of the turntable in the camera coordinate system are obtained as follows.

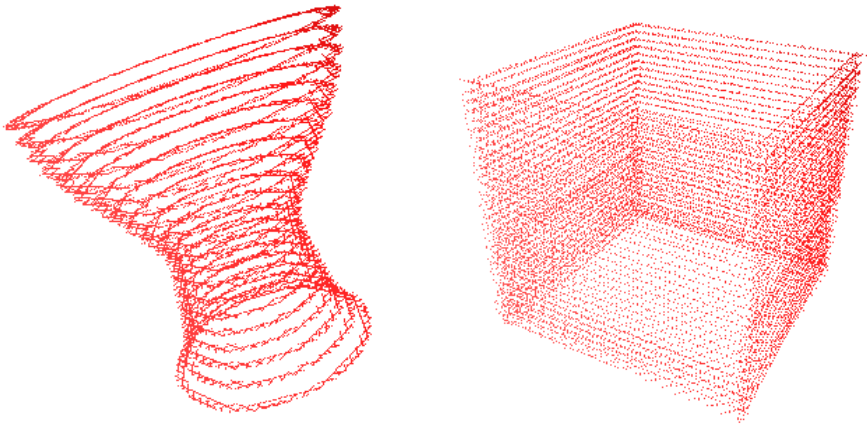
It is well-known that the perspective projection of a circle in the 3D space is an ellipse in the image. This projective mapping is a homography and can be used to determine the relative orientation between the image plane and the plane consisting of the circle. In the implementation, Canny edge detection is first applied on the image with only the turntable [12], followed by a least square fitting algorithm to detect the elliptical shape in the resulting edge image [13].



(a) Visible surface with level curves of the vase object captured by a virtual camera. Only five images out of 20 captures for a single view are shown in the figure.



(b) Visible surface with level curves of the cube object captured by a virtual camera. Only five images out of 30 captures for a single view are shown in the figure.



(c) Reconstructed 3D point clouds of the computer generated objects.

Fig. 2. Simulation results with synthesis data sets

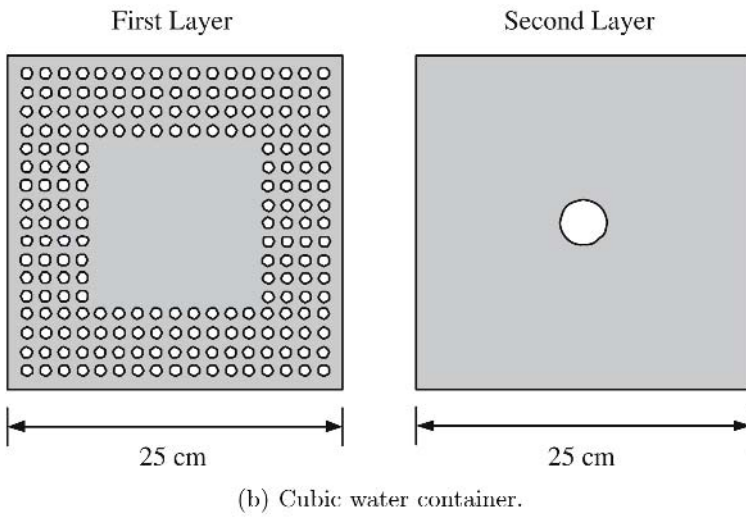
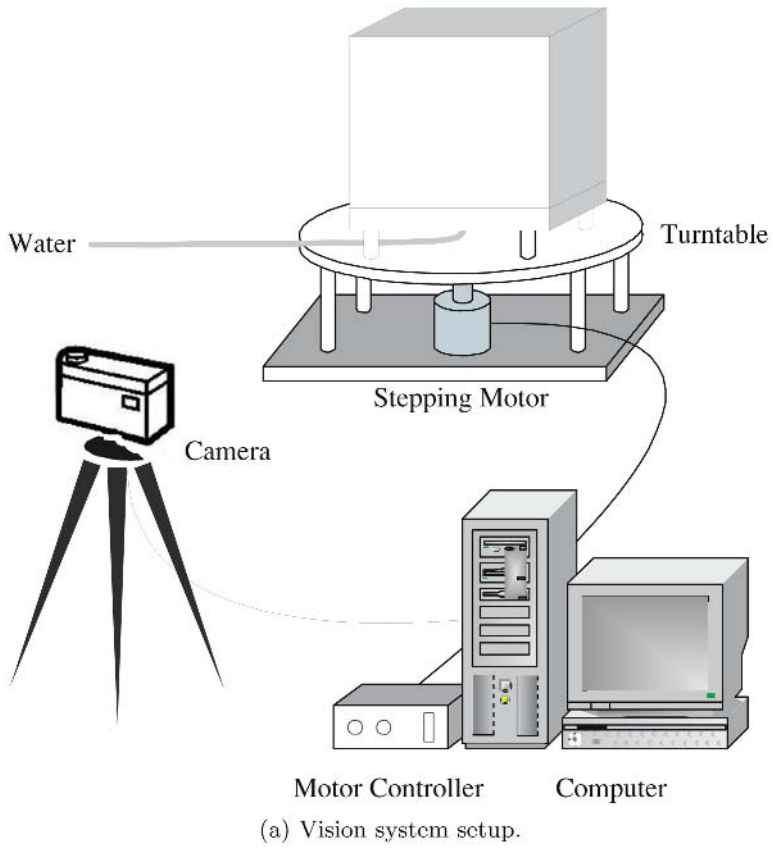


Fig. 3. 3D acquisition system and experimental setup

The rotation matrix between the camera and the turntable is then estimated by an SVD-based pose ellipse algorithm [14]. Since the unit vector of the rotation axis is in the 3D space, its location can be represented by the turntable center. Consequently, the scale factor associated with the perspective projection of the circle is derived from the physical size of the turntable. It is then used to recover the 3D coordinates of the turntable center.

In the experiments, the parallel planes for obtaining the cross-sections of the object are acquired by increasing the water level in the container. The coplanar 3D curves are generated and recorded by a static camera. To obtain the corresponding 2D curve of an object’s cross-section, a quadrilateral image region within the water surface (which is bounded by the cubic container) is extracted and used for edge detection. Since the object might contain some edge features other than its boundaries, cross-section (or level curve) detection is carried out by transforming the quadrilateral region to the HSI color space for region segmentation and boundary extraction. The cross-section curve is then identified by scanning the resulting edge image horizontally from left to right, and searching for the “smooth” edge segment with a predefined orientation change range for two neighboring pixels in the curve (typically from -45° to 45°).

As shown in Section 2, for a given water level the corresponding plane equation can be derived from the image points of the water surface, i.e. the corner points of the surface patch bounded by the cubic container. First, edge detection and Hough transform are applied to identify the line features in the image. The vertical lines which represent the edges of the container are removed from the image. The intersections of the remaining straight lines are then selected as candidate corner points for the plane equation and further checked with the water region segmentation. Only the four points near the boundary of the water surface region are selected for plane parameters computation. Fig. 4 shows the results of several recorded images with coplanar corner point detection and the extracted 2D curves.

5 Experimental Results

We have tested the proposed shape from parallel planar 3D curves algorithm on several real objects. To mitigate the object’s reflection on the water surface, white colored liquid is used. In general, different colors can be selected according to the object’s appearance. For single viewpoint acquisition, the water level is raised continuously during the image captures. But for the multi-view 3D model reconstruction, the water level remains static during the rotation of the turntable (every 90° with four image captures). To make sure the 2D curves distinguishable between the images, only the images with water level difference greater than 10 pixels are used. In the experimental setup, it corresponds to about 5 *mm* in the real scene. The experimental result of an object “chicken” is shown in Fig. 5.

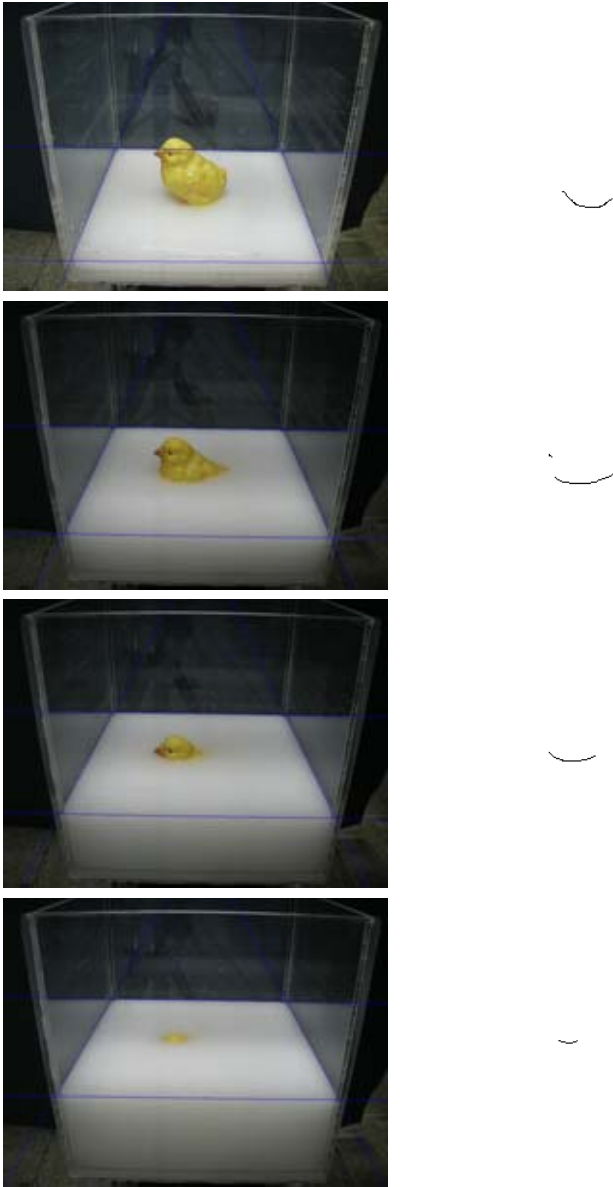


Fig. 4. Coplanar corner points and 2D curves

In addition to the correctness of the camera parameters, the accuracy of the proposed method also depends on the resolution of the images and imprecise level curve formation due to water viscosity. A rudimentary error analysis is carried out using a cylindrical object with known dimension. Figures 6(a) and 6(b) show the results when the camera's viewpoint is perpendicular and with a

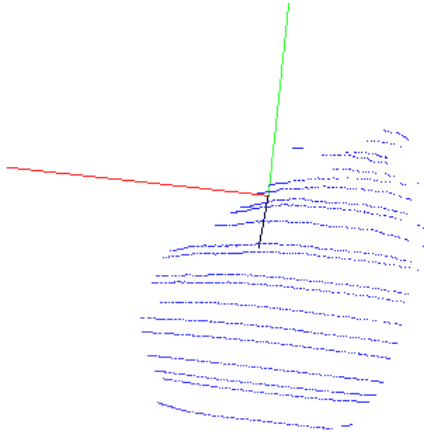


Fig. 5. Experimental Result

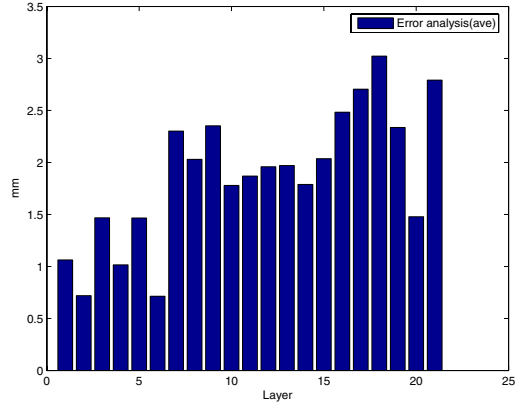
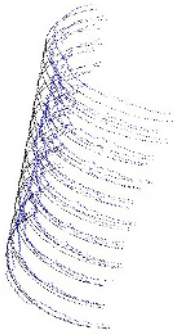
tilt angle to the turntable axis, respectively. The left figures show the real surface points (in black) and the recovered 3D curves (in blue). The average errors in each slice (in *mm*) for both cases are illustrated in the right figures. It can be seen that the error roughly increases with the water level. The reason could be the inaccurate depth computation due to the foreshortening of the surface patch, especially for the second case.

6 Conclusion

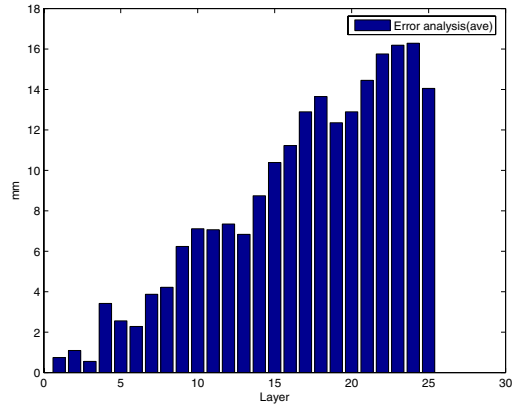
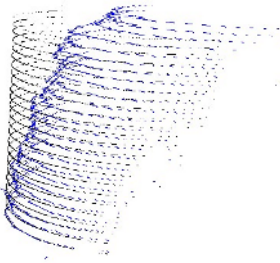
In this paper we present a 3D model acquisition system based on projections of the object's level curves, and demonstrate the results from a prototype implementation. The proposed method is similar to 3D reconstruction from CT or laser range scanning, but our approach can be implemented on a low-cost passive camera system. The system parameters can be obtained by self-calibration without elaborate experimental setup. Since the 3D coordinates are computed directly from the corresponding image points, dense depth map can only be achieved by high resolution 2D curves. Currently the average error is less than 3 *mm* in the working range of about 1 meter from the camera. In the future work, partial 3D shapes acquired from multiple viewpoints will be integrated to create a complete 3D model.

Acknowledgments

The support of this work in part by the National Science Council of Taiwan, R.O.C. under Grant NSC-94-2213-E-194-041 is gratefully acknowledged.



(a) The camera's viewpoint perpendicular to the turntable axis.



(b) The camera's viewpoint is adjusted with a tilt angle.

Fig. 6. Error analysis

References

1. Forsyth, D., Ponce, J.: Computer Vision: A Modern Approach. Prentice-Hall (2003)
2. Lin, H.Y., Subbarao, M.: A vision system for fast 3d model reconstruction. In: CVPR (2), IEEE Computer Society (2001) 663–668
3. Nayar, S., Nakagawa, Y.: Shape from focus. IEEE Trans. Pattern Analysis and Machine Intelligence **16** (1994) 824–831
4. Baba, M., Asada, N., Oda, A., Migita, T.: A thin lens based camera model for depth estimation from blur and translation by zooming. In: Vision Interface. (2002) 274

5. Fuchs, H., Kedem, Z.M., Useton, S.P.: Optimal surface reconstruction from planar contours. *Commun. ACM* **20** (1977) 693–702
6. Barequet, G., Shapiro, D., Tal, A.: Multilevel sensitive reconstruction of polyhedral surfaces from parallel slices. *The Visual Computer* **16** (2000) 116–133
7. Meyers, D., Skinner, S., Sloan, K.R.: Surfaces from contours. *ACM Trans. Graph.* **11** (1992) 228–258
8. Klein, R., Schilling, A., Straßer, W.: Reconstruction and simplification of surfaces from contours. *Graphical Models* **62** (2000) 429–443
9. Lin, H.Y.: Vehicle speed detection and identification from a single motion blurred image. In: *WACV/MOTION*, IEEE Computer Society (2005) 461–467
10. Chen, C., Yu, C., Hung, Y.: New calibration-free approach for augmented reality based on parameterized cuboid structure. In: *International Conference on Computer Vision*. (1999) 30–37
11. Schroeder, W., Martin, K.M., Lorensen, W.E.: *The visualization toolkit* (2nd ed.): an object-oriented approach to 3D graphics. Prentice-Hall, Inc., Upper Saddle River, NJ, USA (1998)
12. Canny, J.: A computational approach to edge detection. *IEEE Trans. Pattern Analysis and Machine Intelligence* **8** (1986) 679–698
13. Fitzgibbon, A.W., Pilu, M., Fisher, R.B.: Direct least square fitting of ellipses. *IEEE Trans. Pattern Anal. Mach. Intell.* **21** (1999) 476–480
14. Trucco, E., Verri, A.: *Introductory Techniques for 3-D Computer Vision*. Prentice Hall (1998)

Scale Invariant Robust Registration of 3D-Point Data and a Triangle Mesh by Global Optimization

Onay Urfaloğlu, Patrick Mikulastik, and Ivo Stegmann

Information Technology Laboratory (LFI), University of Hannover

Abstract. A robust registration of 3D-point data and a triangle mesh of the corresponding 3D-structure is presented, where the acquired 3D-point data may be noisy, may include outliers and may have wrong scale. Furthermore, in this approach it is not required to have a good initial match so the 3D-point cloud and the according triangle mesh may be loosely positioned in space. An additional advantage is that no correspondences have to exist between the 3D-points and the triangle mesh. The problem is solved utilizing a robust cost function in combination with an evolutionary global optimizer as shown in synthetic and real data experiments.

Keywords: scale invariant, robust registration, evolutionary optimization, 3D-transformation.

1 Introduction

Registration is applied in object recognition, 3D-geometry processing and acquisition. Often two observations of a 3D-scene are provided, where one is called the model and the other is called the data. Each of them is defined in its own coordinate system. The process of registration is to find a transformation which best fits the data to its corresponding model. The generation of 3D-point cloud data in the process of monocular camera parameter estimation cannot guarantee a correct scale with respect to a 3D-model, unless the scale is previously determined and applied.

In this paper robust registration of a 3D-point cloud with respect to a provided 3D-model, e.g. a polygon set is addressed. The most common methods for this purpose are based on the Iterated Closest Point (ICP) algorithm [1, 2, 3, 4]. Scale invariant versions of ICP-based methods were realized by [5]. However, the convergence of ICP-based methods to the global optimum requires the data set to be sufficiently prealigned with respect to the corresponding 3D-model. To overcome this problem, another approach was proposed [6] utilizing global optimization in the pose space to find the transformation which best aligns the provided range images. But this approach does not enable the estimation of the scale factor.

The approach proposed in this paper enables automatic registration where no correspondences between the 3D-point cloud and the 3D-model are required.

Similar to the approach in [6], the 3D-point cloud may include outliers and may be freely initially positioned relative to the corresponding 3D-model. However, in contrast to [6], the 3D-point cloud may also have a wrong scale. A modified version of the robust cost function [7] is utilized in combination with an efficient global optimization method called *Differential Evolution* (DE) [8] in order to estimate the correct transformation of the 3D-point cloud.

The paper is organized as follows. In section 2, the modified robust cost function is presented. In section 3, the utilized evolutionary global optimization is described. In the following section 4 experimental results are presented and in the last section 5 the paper is concluded.

2 Robust Cost Function

The proposed cost function is based on the distances of the 3D-points to the triangle mesh. In order to determine the distance of a 3D-point to the triangle mesh it is required to find the closest triangle first. This is accomplished by calculating the distances to all planes defined by each triangle, respectively. Figure 1 shows a case where the 3D-point P has a distance d_0 to the selected

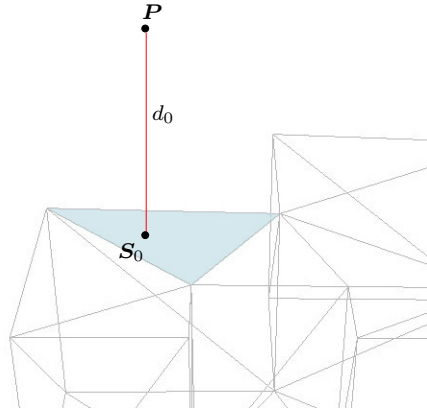


Fig. 1. Distance d_0 to plane defined by the triangle

triangle. In this case it is sufficient to calculate the distance to the plane. Only if the intersection point S_0 in the plane is within the triangle the calculated distance is valid. This is checked by the calculation of the so called Barycentric coordinates. Figure 2 shows a triangle and a Point S_0 , where its corresponding Barycentric coordinates are $(\alpha_1, \alpha_2, \alpha_3)$ and W_i are the vectors pointing to the vertices of the triangle

$$S_0 = \alpha_1 W_1 + \alpha_2 W_2 + \alpha_3 W_3. \quad (1)$$

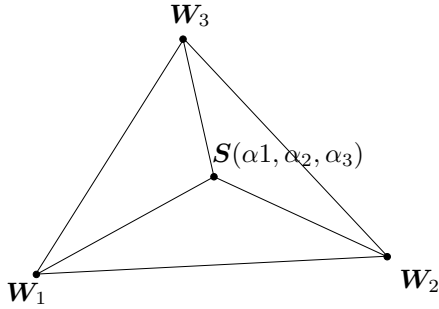


Fig. 2. Barycentric coordinates used to test whether a point S_0 is within the triangle

The point S_0 is within the triangle if and only if

$$\alpha_1, \alpha_2, \alpha_3 \in [0, 1] \wedge \alpha_1 + \alpha_2 + \alpha_3 = 1. \tag{2}$$

In contrast, figure 3 shows a case where the intersection point S_0 is not within the triangle. Therefore, in this case all 3 distances d_1, d_2, d_3 to the lines defined

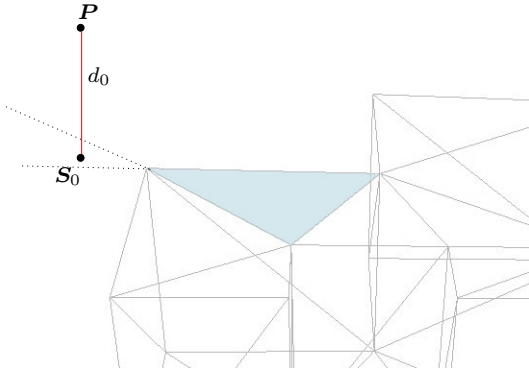


Fig. 3. Case where the intersection Point S_0 is not within the triangle

by the edges of the triangle have to be determined additionally, as shown in figure 4.

But even the determination of these distances proves to be insufficient, since the closest point on a line is not necessarily within the triangle, so three additional distances d_4, d_5, d_6 to the vertices have to be calculated, as shown in figure 5. Finally, the smallest distance d having the corresponding intersection point S_i out of all calculated distances $d_i, i \in [0, 6]$ is chosen.

The distances depend on the orientation and the position of the 3D-point cloud. The task is to estimate the transformation consisting of the orientation,

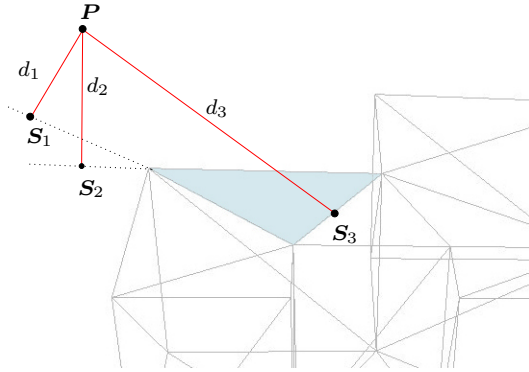


Fig. 4. Distances d_1, d_2, d_3 to lines defined by the edges of the triangle

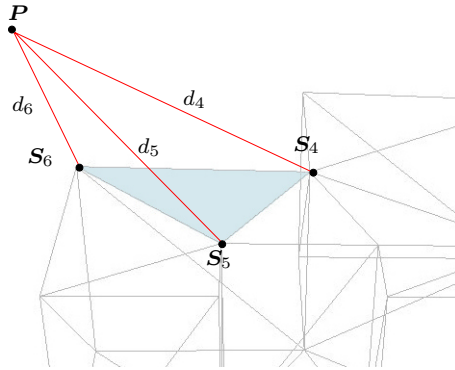


Fig. 5. Distances d_4, d_5, d_6 to vertices of the triangle

the translation and the overall scale factor of the 3D-point cloud which leads to a best match with the triangle mesh. The rotation matrix \mathbf{R} is parameterized by the angles $\phi, \theta, \rho \in [0, \pi]$

$$\mathbf{R}(\phi, \theta, \rho). \tag{3}$$

The translation vector \mathbf{t} is denoted

$$\mathbf{t} = (\delta x, \delta y, \delta z). \tag{4}$$

The overall scale factor is denoted s . The parameter vector \mathbf{x} is defined by

$$\mathbf{x} = (\phi, \theta, \rho, \delta x, \delta y, \delta z, s)^\top \tag{5}$$

and contains all 7 parameters to be estimated. A 3D-point \mathbf{P} is then transformed to \mathbf{P}' by the transformation \mathbf{T}

$$\mathbf{P}' = \mathbf{T}(\mathbf{x})\mathbf{P} = s(\mathbf{R}(\phi, \theta, \rho)\mathbf{P} + \mathbf{t}). \tag{6}$$

Therefore, each distance d_j of the point \mathbf{P}_j depends on the parameter vector and is determined by

$$d_j = d_j(\mathbf{x}) = |\mathbf{T}(\mathbf{x})\mathbf{P}_j - \mathbf{S}|. \tag{7}$$

Since the 3D-point cloud consisting of N points may include outlier points, a robust cost function $\Gamma(\mathbf{x})$ is utilized, which is a modified version of the cost function defined in [7]. Assuming a Gaussian error with zero mean and a variance of σ for the position of the point \mathbf{P} , the utilized cost function is defined by

$$\Gamma(\mathbf{x}) = \sum_{j=1}^N -\exp\left(-\frac{d_j(\mathbf{x})^2}{2\kappa s^3}\right), \tag{8}$$

where the parameter κ depends mainly on σ . As a rule of thumb, good estimates are obtained for

$$\kappa \approx 10\sigma^2. \tag{9}$$

This cost function has to be minimized in order to find the best solution vector \mathbf{x} .

Generally, this robust cost function results in local minima in the search space where the *global* minimum has to be found. The number of local minima increases with the number of outliers. The scale as an additional degree of freedom leads to even more complicated search spaces. Furthermore, the initial orientation and the position of the 3D-point cloud may differ considerably with respect to the triangle mesh, so it is appropriate to use a global optimization method, which is described in the next section.

The modification of the robust const function is realized by the multiplicative term s^3 in the denominator of (8) which enforces an additional weighting of the distance depending on the scale. The reason for this modification is that there is always an attraction towards smaller scale factors since, in the average, this leads to smaller distances. This property is (over) compensated by the introduction of this term in order to enhance the global search in greater scale factor regions.

3 Differential Evolution Optimization

The utilized evolutionary optimizer is the so called *Differential Evolution* (DE) method [8, 9, 10]. It is known as an efficient global optimization method for continuous problem spaces with many applications. The optimization is based on a population of $n = 1, \dots, M$ solution candidates $\mathbf{x}_{n,i}$ at iteration i where each candidate has a position in the 7-dimensional search space. Initially, the solution candidates are randomly generated within the provided intervals of the search space. The population improves by generating new positions iteratively for each candidate. New positions for the iteration step $i + 1$ are determined by

$$\mathbf{y}_{n,i+1} = \mathbf{x}_{k,i} + F \cdot (\mathbf{x}_{l,i} - \mathbf{x}_{m,i}) \tag{10}$$

$$\mathbf{x}_{n,i+1} = C(\mathbf{x}_{n,i}, \mathbf{y}_{n,i+1}), \tag{11}$$

where k, l, m are random integers from interval $[1, M]$, F is a weighting scalar, $\mathbf{y}_{n,i+1}$ a displaced $\mathbf{x}_{k,i}$ by a weighted difference vector and $C()$ is a crossover operator copying coordinates from both $\mathbf{x}_{n,i}$ and $\mathbf{y}_{n,i+1}$ in order to create $\mathbf{x}_{n,i+1}$. The crossover operator C is provided with a value specifying the probability to copy coordinates either from $\mathbf{x}_{n,i}$ or $\mathbf{y}_{n,i+1}$ to $\mathbf{x}_{n,i+1}$. Only if the new candidate $\mathbf{x}_{n,i+1}$ proves to have a lower cost it replaces $\mathbf{x}_{n,i}$, otherwise it is discarded.

DE includes an adaptive range scaling for the generation of solution candidates through the difference term in (10). This enables global search in the case where the solution candidate vectors are spread in the search space and the mean difference vector is relatively large. In the case of a converging population the mean difference vector becomes relatively small and this enables efficient fine tuning at the end phase of the optimization process.

4 Experimental Results

In order to test the proposed approach both synthetic and real data based experiments are performed.

4.1 Synthetic Data

As shown in fig. 6, 4 boxes represented by their wireframes are placed in 3D-space. The corresponding 3D-point cloud consists of the corner points of the boxes with no position error. Additionally, the 3D-point cloud includes 4 outlier points indicated by surrounding squares. For all points \mathbf{P}_j , it is

$$\mathbf{P}_j \in [-6, 12]u \times [0, 21]u \times [0, 41]u, \quad (12)$$

where u is any arbitrary length unit. By transforming and rescaling the coordinates of the 3D-point cloud a second 3D-point cloud is generated and the robust registration is performed utilizing the second point cloud. The 3D-point cloud is translated by $(0 \ 5 \ 5)^\top u$ and rotated by $\phi = 0.4$ within the x-y-plane. 3 different global scales $s = 0.5, 1, 2$ are applied. The search interval for the translation coordinates is set to $[-20, 20]$, whereas the search interval or the scale factor is set to $[0.5, 2]$. The population size of the DE-algorithm is set to 60.

Fig. 7 shows the boxes and the corresponding transformed 3D-point cloud with a scale factor of $s = 0.5$.

Fig. 8 and 9 show the same configuration with a scale factor of $s = 1$ and $s = 2$, respectively.

In all experiments with synthetic data our algorithm found the global minimum of the cost function, so that the correct transformation and scale parameters could be recovered, as shown in fig. 6.

4.2 Real Data

In order to test the proposed approach in real data based scenarios, two real world 3D-objects are modelled by hand and with a 3D scanner to create the

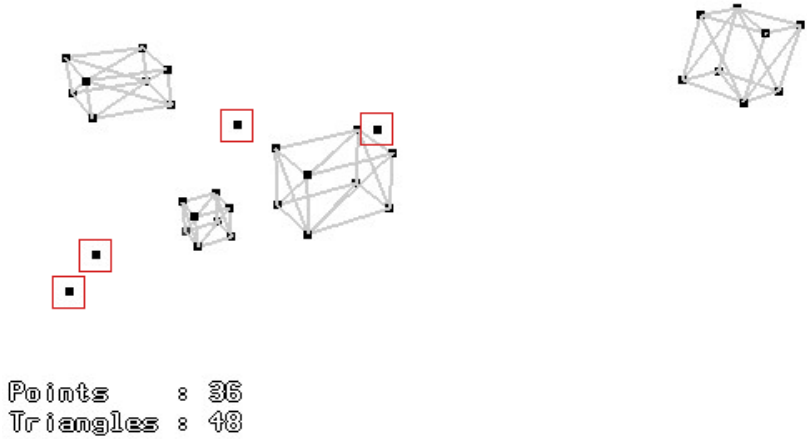


Fig. 6. Result of the robust registration

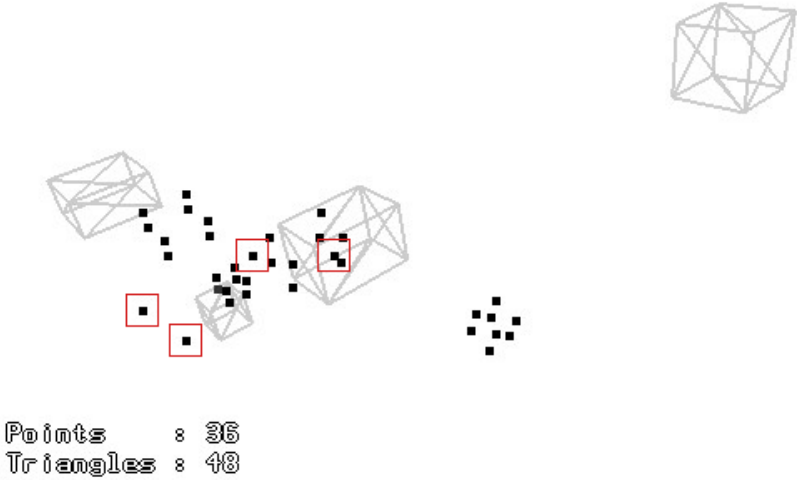


Fig. 7. Initial configuration of 3d-point cloud and corresponding triangle mesh at $s = 0.5$

corresponding triangle meshes. Fig.10 shows the 3D-objects named 'block' and 'dino', respectively. In the following experiments, the search interval for the coordinates of the translation vector is set to $[-300, 300]$. The search interval for the scale factor is set to $[0.5, 2]$. The corresponding 3D-point cloud is generated by structure-from-motion utilizing a monocular camera. In both cases, the average error variance of the 3D-points of the resulting 3D-clouds is $\sigma^2 \approx 2.25u^2$, assuming a Gaussian probability density for the position error. Figure 11 shows the initial configuration of the test object 'block'.

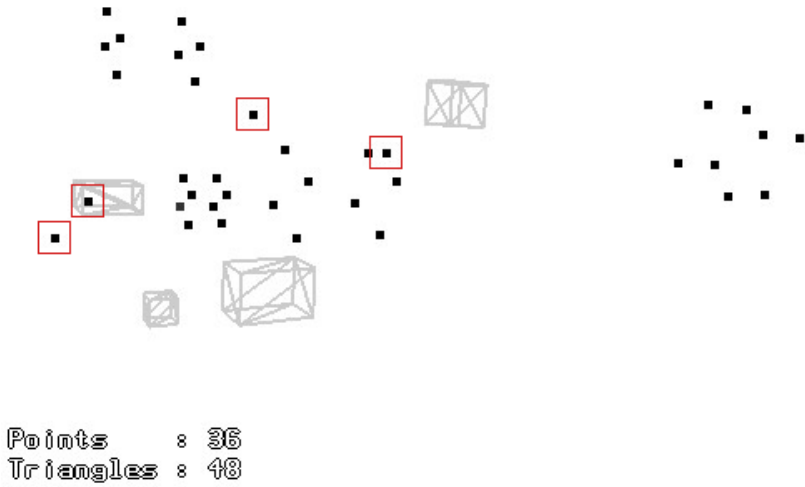


Fig. 8. Initial configuration of 3d-point cloud and corresponding triangle mesh at $s = 1$

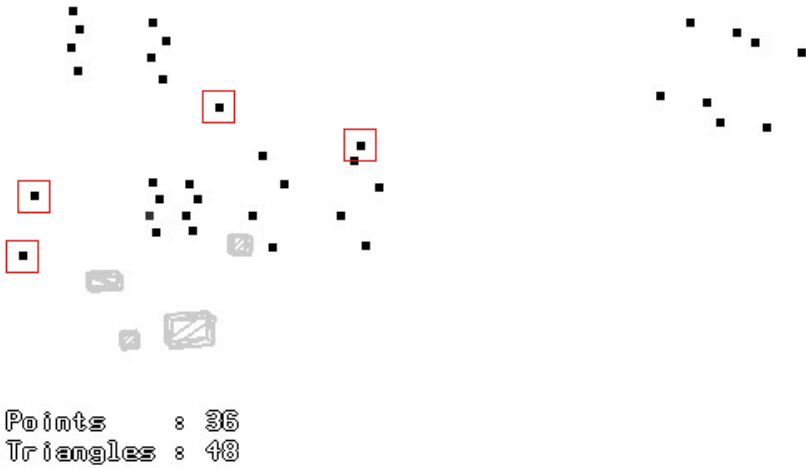


Fig. 9. Initial configuration of 3d-point cloud and corresponding triangle mesh at $s = 2$

The corresponding 3D-point cloud consists of $N = 1103$ points \mathbf{P}_j where

$$\mathbf{P}_j \in [-125, 120]u \times [-132, 105]u \times [-25, 80]u. \quad (13)$$

The corresponding triangle mesh is composed of 56 triangles. Figure 12 shows the result of the robust registration, achieved after 500 iterations at a population size of 60. The estimated required scale factor is $s = 0.514$, the number of outliers

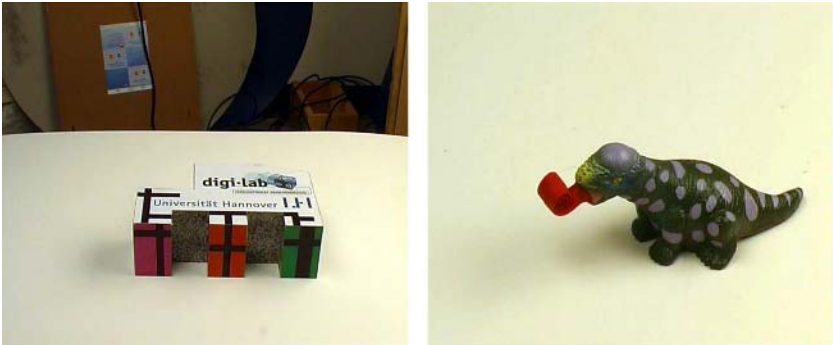


Fig. 10. 3D-objects used for real data experiments

is 214 at an inlier threshold of $\sigma_i = 3u$. The std. deviation of the inlier-error is $\sigma_e = 1.219u$.

In the second real data test for the proposed approach, the initial configuration of the 3D-point cloud with respect to the corresponding triangle mesh is shown in figure 13. The point cloud consists of $N = 904$ points P_j where

$$P_j \in [-125, 120]u \times [-132, 105]u \times [-25, 80]u. \tag{14}$$

The corresponding triangle mesh is composed of 2000 triangles. The result of the robust registration is shown in figure 14, achieved after 900 iterations at

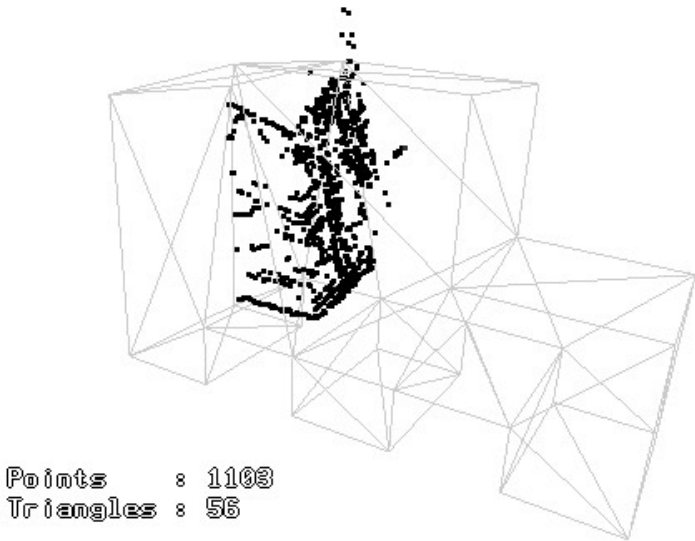


Fig. 11. 'block': Initial configuration of 3D-point cloud and the corresponding triangle mesh

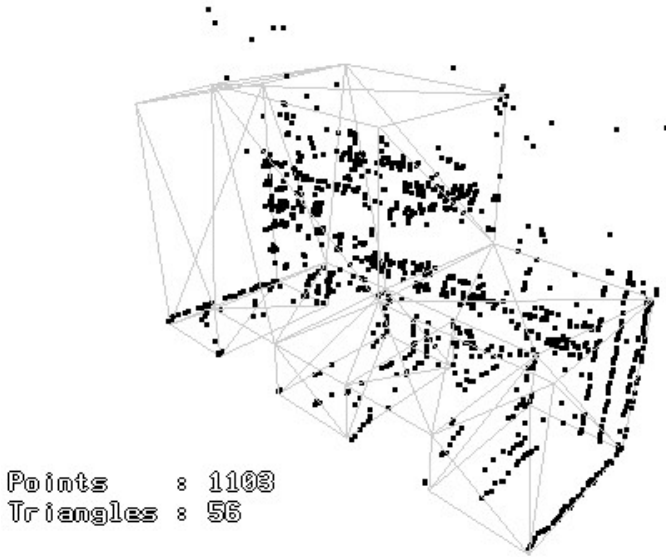


Fig. 12. 'block': Result of the registration of the 3D-point cloud

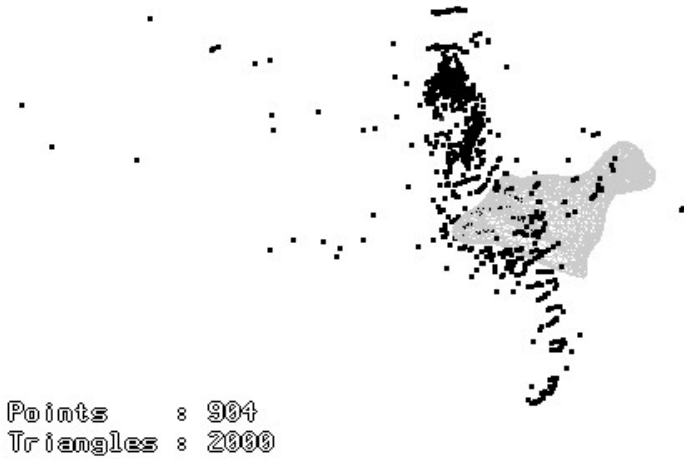


Fig. 13. 'dino': Initial configuration of 3D-point cloud and the corresponding triangle mesh

a population size of 90. The estimated required scale factor is $s = 1.874$, the number of outliers is 52 at an inlier threshold of $\sigma_i = 3u$. The std. deviation of the inlier-error is $\sigma_e = 0.687u$.

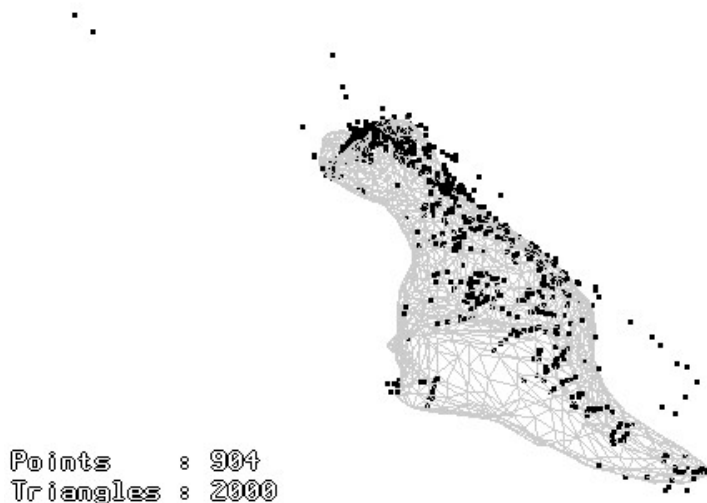


Fig. 14. 'dino': Result of the registration of the 3D-point cloud

5 Conclusions

The proposed approach enables the robust registration of two observations of a 3D-object, where one is represented as a triangle mesh and the other is a 3D-point cloud, in different coordinate systems at different scales. With this approach the automatic and accurate estimation of the transformation parameters from one system to the other is made possible, even in cases of high outlier amounts and lack of prealignment of the two coordinate systems. The additional scale parameter proved to further complicate the estimation, so the utilized cost function had to be adapted accordingly. This approach is applicable in cases where the registration has to be done automatically without manual prealignment and where the scale factors of the two coordinate systems may be different.

References

1. Besl, P., McKay, N.: A method for registering of 3-d shapes. In: PAMI. Volume 14. (1992) 239–256
2. Chen, C.S., Hung, Y.P., Cheung, J.B.: Ransac-based darces: a new approach to fast automatic registration of partially overlapping range images. In: IEEE Transactions on Pattern Analysis and Machine Intelligence. Volume 21. (1999) 1229–1234
3. Masuda, T., Yokoya, N.: A robust method for registration and segmentation of multiple range images. *Computer Vision and Image Understanding* **61** (1995) 295–307

4. Sharp, G.C., Lee, S.W., Wehe, D.K.: Invariant features and the registration of rigid bodies. In: IEEE International Conference on Robotics & Automation. (1999) 932–937
5. Burschka, D., Li, M., Taylor, R., Hager, G.D.: Scale-invariant registration of monocular stereo images to 3d surface models. In: ICIRS, Sendai, Japan (2004) 2581–2586
6. Silva, L., Bellon, O.R.P., Boyer, K.L.: Robust multiview range image registration. In: Proc. of the XVI Brazilian Symposium on Computer Graphics and Image Processing. (2003)
7. Urfahoglu, O.: Robust estimation with non linear particle swarm optimization. In: Proceedings of Mirage 2005 (Computer Vision / Computer Graphics Collaboration Techniques and Applications), INRIA Rocquencourt, France (2005)
8. Storn, R., Price, K.: Differential evolution - a simple and efficient adaptive scheme for global optimization over continuous spaces. Technical Report TR-95-012, ICSI (1995)
9. Storn, R., Price, K.: Minimizing the real functions of the icec'96 contest by differential evolution. In: IEEE International Conference on Evolutionary Computation, Nagoya (1996) 842–844
10. Price, K.V.: Differential evolution: a fast and simple numerical optimizer. In: Biennial Conference of the North American Fuzzy Information Processing Society, NAFIPS, IEEE Press, New York. ISBN: 0-7803-3225-3 (1996) 524–527

Fast Hough Transform Based on 3D Image Space Division

Witold Zorski

Cybernetics Faculty, Military University of Technology
S. Kaliskiego 2, 00-908 Warsaw, Poland
wzorski@ita.wat.edu.pl

Abstract. This paper presents a problem of 3D images decomposition into spheres. The presented method is based on a fast Hough transform with an input image space division. An essential element of this method is the use of a clustering technique for partial data sets. The method simplifies the application of Hough transform to segmentation tasks as well as accelerates calculations considerably.

1 Introduction to the Hough Transform

The Hough transform was patented in 1962 as a method for detecting complex patterns of points in binary images [3]. In 1981 Deans noticed [2] that the Hough transform for straight lines was a specific case of the more general Radon transform [8] known since 1917, which is defined as (for function $I(x, y)$ in two-dimensional Euclidean space):

$$H(\rho, \alpha) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(x, y) \delta(\rho - x \cos(\alpha) - y \sin(\alpha)) dx dy \quad (1)$$

where δ is the delta function. This formula shows that the function $I(x, y)$ is integrated along the straight line determined by the parametric equation $\rho = x \cos(\alpha) + y \sin(\alpha)$. The Radon transform is equivalent to the Hough transform when considering binary images (i.e. when the function $I(x, y)$ takes values 0 or 1). The Radon transform for shapes other than straight lines can be obtained by replacing the delta function argument by a function, which forces integration of the image along contours appropriate to the shape.

Using the Radon transform to calculate the Hough transform is simple (almost intuitive) and is often applied in computer implementations. We call this operation **pixel counting** in the binary image.

An (alternative) interpretation of the Hough transform is the so-called **backprojection** method. The detection of analytical curves defined in a parametrical way, other than straight lines is quite obvious. Points (x, y) of image lying on the curved line determined by n parameters a_1, \dots, a_n may be presented in the form:

$$\lambda_o = \{(x, y) \in \mathbb{R}^2 : g(\hat{a}_1, \dots, \hat{a}_n, (x, y)) = 0\} \tag{2}$$

where $g(\hat{a}_1, \dots, \hat{a}_n, (x, y)) = 0$ describes the given curve.

By exchanging the meaning of parameters and variables in the above equation we obtain the backprojection relation (mapping image points into parameter space), which may be written down in the following way:

$$\lambda_T = \{(a_1, \dots, a_n) \in \mathbb{R}^n : g(\hat{x}, \hat{y}, (a_1, \dots, a_n)) = 0\} \tag{3}$$

Based on (3) the Hough transform $H(a_1, \dots, a_n)$ for image $I(x, y)$ is defined as follows:

$$H(a_1, \dots, a_n) = \sum_{(x_i, y_i) \in I} h(\hat{x}_i, \hat{y}_i, a_1, \dots, a_n) \tag{4}$$

where

$$h(\hat{x}_i, \hat{y}_i, a_1, \dots, a_n) = \begin{cases} 1 & \text{if } g(\hat{x}_i, \hat{y}_i, (a_1, \dots, a_n)) = 0 \\ 0 & \text{otherwise.} \end{cases} \tag{5}$$

In order to calculate the Hough transform digitally an appropriate representation of the parameter space $H(a_1, \dots, a_n)$ is required. In a standard implementation, any dimension in the parameter space is subject to quantisation and narrowing to an appropriate range. As a result, an array is obtained where any element is identified by the parameters (a_1, \dots, a_n) . An element in the array is increased by 1 when the analytical curve, determined by co-ordinates (a_1, \dots, a_n) , passes through point (\hat{x}, \hat{y}) of the object in image I . This process is called **accumulation** and the array used is called an **accumulator** (usually marked with symbol A).

Thus, we may assume that the Hough transform is based on a representation of the image I into the accumulator array A , which is defined as follows:

$$A : P \rightarrow \mathbb{N}, \quad \text{where } P = P_1 \times P_2 \times \dots \times P_p. \tag{6}$$

The symbol $P_i \subset \mathbb{N}$ determines the range of i -parameters of a p -dimensional space P . Determining array A is conducted through the calculation of partial values for points of an object in image I and adding them to the previous ones (see 4) which constitutes the process of accumulation. Initially, all elements of array A are zeros.

This paper presents an application of the Hough transform to the tasks of identifying spheres in 3D images. Although spheres are described by four parameters only **one-dimensional accumulator array will be used** in the presented method. An important element of this method is the use of a clustering technique for partial results. The method simplifies the application of Hough transform to segmentation tasks as well as accelerates calculations considerably.

Lots of works [1], [7] have been done in the area of fast Hough transform since [6]. The main difference is that **in the presented method the space of an input image is**

divided into fragments what is called “image space division”. The same approach was used by the author in the case of straight lines [10] and circles [13].

2 The Fast Hough Transform – General Approach [6]

The main assumption of so-called fast Hough transform is that the input image space features “vote” for sets of points, lying on hyper planes in the parameter space. This recursively divides the parameter space into hyper cubes from low to high resolution and performs the Hough transform only on the hyper cubes with votes exceeding a selected threshold. In the case of two and three dimensional parameter space these hyper cubes are respectively square sectors or cubic sectors of the parameter space.

In the case of the “standard” Hough transform a parameter space is treated as the “array of accumulators” and points of concentration are found by identifying those accumulators receiving the largest number of votes. In that formulation, both the computational complexity of the voting process and the memory for the votes grow exponentially with the quantisation and the dimensionality of the parameter space. Therefore the standard Hough transform is not feasible for tasks with high resolution or high dimensionality which is exactly the case of spheres.

3 The Circle Hough Transform

As the starting point for spheres identification in 3D images let us look first at a simpler and analogous task i.e. circles identification in 2D images. The problem is described in details by the author in his earlier paper [13]. Nevertheless the most important thing is to catch the procedure used to solve the problem.

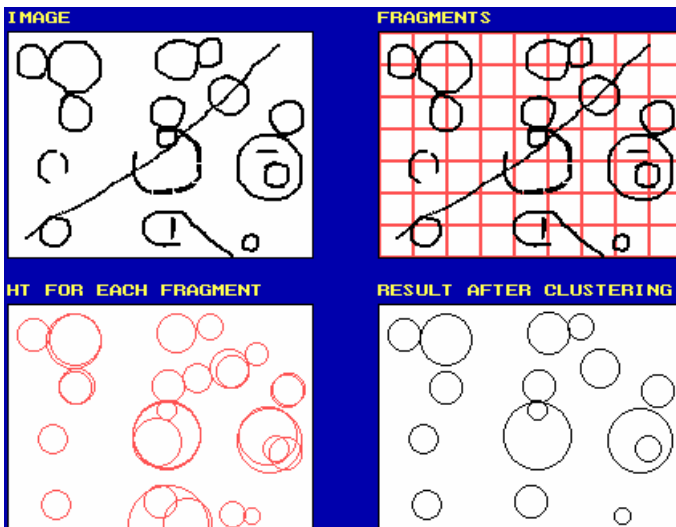


Fig. 1. An example result for the circle Hough transform

The method is presented in Fig. 1 and consists of partitioning the image space into fragments and using Hough transform for each of them (one accumulator table A is identified for each image fragment). Then, global maxima are found in every table A. Finally, a clustering technique is used.

It is necessary mention that the considered method is very similar to the one described by the author in [10] which concerns straight lines detection.

4 Basic Definitions

This paper considers **3D digital binary images**, i.e. three-dimensional images, which are formed with sets of points, by convention either black or white. Such an image (binary image) may be presented as the following function:

$$I : D \rightarrow \{0,1\}, \text{ where } D = [1, \dots, K] \times [1, \dots, L] \times [1, \dots, M] \subset N^3. \tag{7}$$

Hence, we may consider a 3D image as a cubic matrix of which row, column and “deep” indices identify a pixel of the image.

Given an image I, we can generally define an **object** b(I) as follows:

$$b(I) = \{(k, l, m) \in D : I(k, l, m) = 1\}. \tag{8}$$

Let us denote every image that originated from image I as the result of restricting its domain D, as a fragment Q of image I. Thus, a **fragment** Q of image I is defined as follows:

$$Q : D_Q \rightarrow \{0,1\}, \text{ where } D_Q \subset D = [1, \dots, K] \times [1, \dots, L] \times [1, \dots, M] \subset N^3. \tag{9}$$

The Hough transform $H(x_s, y_s, z_s, r)$ for spheres detection in 3D images may be given by

$$H(x_s, y_s, z_s, r) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(x, y, z) \delta((x - x_s)^2 + (y - y_s)^2 + (z - z_s)^2 - r^2) dx dy dz \tag{10}$$

or

$$H(x_s, y_s, z_s, r) = \int_{-\pi/2}^{\pi/2} \int_{-\pi}^{\pi} I(x_s + r \cos(\alpha) \cos(\beta), y_s + r \sin(\alpha) \cos(\beta), z_s + r \sin(\beta)) d\alpha d\beta, \tag{11}$$

and the parameter space is determined as follows:

$$P = P_1 \times P_2 \times P_3 \times P_4 = [1, \dots, K] \times [1, \dots, L] \times [1, \dots, M] \times [R_{\min}, \dots, R_{\max}] \tag{12}$$

5 The Workshop Issue

In order to conduct some experiment in the case of 3D images it was necessary to build an appropriate tool. The tool was written in Delphi and it is a program. Using

this tool it is possible to generate a set of spheres in a space of size 250x250x250 pixels. Every generated sphere is determined independently by its centre coordinates and its radius and the whole process is based on parametric equations used in formula (11). Additionally the surface of every sphere may be generated according to a specific filling factor and rough factor. It is possible to generate spheres that criss-cross or include one another. An example of 3D input image is shown in Fig. 2.

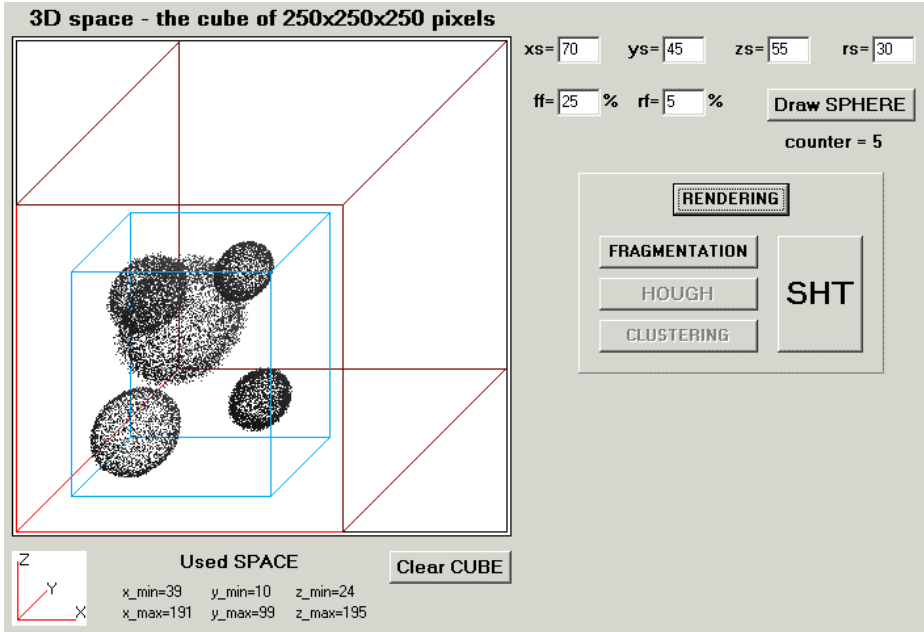


Fig. 2. The tool and example 3D image

6 The Fast Hough Transform with Space Division

This section describes the fast Hough transform as a method of identifying spheres in 3D images.

At the beginning let us consider so called standard approach (see 10 or 11) and its disadvantages. In order to perform segmentation of the object $b(I)$ by means of the standard Hough transform for spheres, the table A must be calculated for the given image I. In practice this becomes computationally difficult for the following reasons:

- Memory requirements. For example an image of size 250x250x250 pixels would require 4D accumulator having more than 10^9 elements.
- Computational complexity. Because the number of parameters is proportional to the number of calculations.
- Problem of spotting of local maxima in the obtained table A. There are many spurious maxima not indicative of continuous spheres in a given image.

The proposed method consists of partitioning the 3D input image into cubic fragments and computing Hough transform for each of them (a 1D table A is identified for each image fragment). Then, global maxima are found in every table A. Finally, a clustering technique is used.

The first step of the suggested method is the partitioning of the image into q fragments. Fig. 3 shows the partitioning results for two examples.

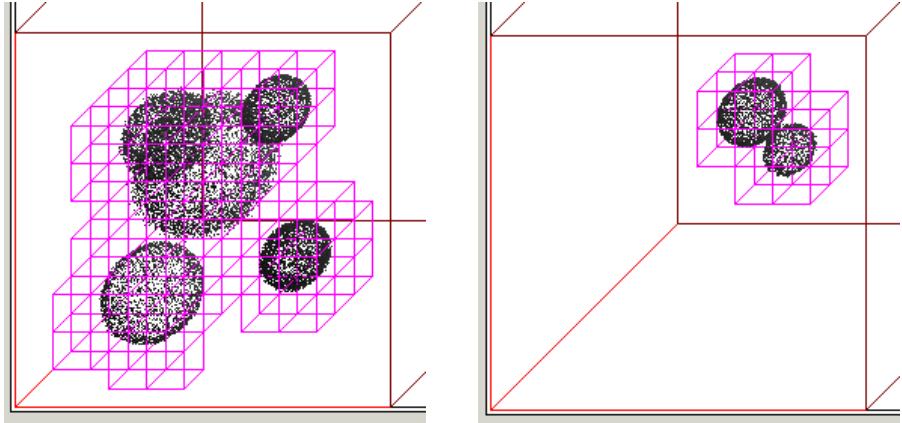


Fig. 3. Two examples of image partitioning (number q of fragments: **116** and **14**)

The partitioning radically influences the calculation rate of the Hough transform. The smaller the fragments are, the lower the calculation complexity is. On the other hand, the reduction of the fragment size may have negative influence on the results of segmentation.

The next step is to determine the 1D accumulator A (i.e. radial histogram):

$$A_i^{(x,y,z)} : P_R \rightarrow N, \text{ where } P_R = P_4 = [R_{\min}, \dots, R_{\max}], \tag{13}$$

for each pixel $(x, y, z) \in I$ **and for each fragment** Q_i **of the image** I (where $i = 1, \dots, q$). That is, for every pixel $(x_Q, y_Q, z_Q) \in Q_i$ we compute its distance from the potential sphere's centre $(x, y, z) \in I$:

$$r = \text{int} \left(\sqrt{(x - x_Q)^2 + (y - y_Q)^2 + (z - z_Q)^2} \right) \tag{14}$$

Then, if $R_{\min} \leq r \leq R_{\max}$, we increment the accumulator location $A_i^{(x,y,z)}[r]$. The maxima of the histograms $A_i^{(x,y,z)}$ correspond to the radii of possible spheres for Q_i . **It is necessary to remember only the best location** $(x_i, y_i, z_i) \in I$ **and radius** r_i **for the maximum value within the array** $A_i^{(x,y,z)}$. This gives us for each fragment Q_i a quadruplet $t_i = (x_i, y_i, z_i, r_i) \in P$. The set of quadruplets $T = \{t_1, \dots, t_q\}$ obtained through calculation accumulators describes the best sphere for each fragment.

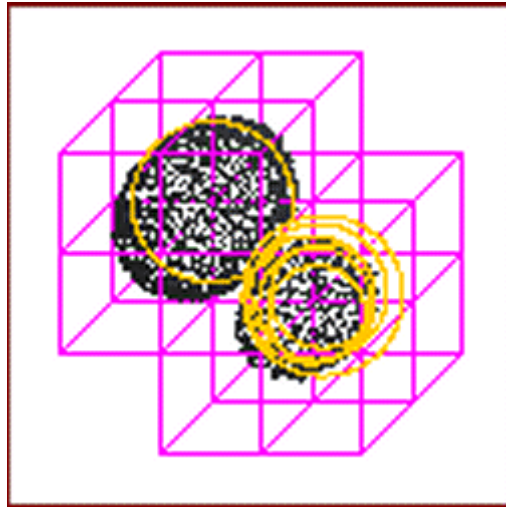


Fig. 4. The partial result obtained for the second example (see Fig. 3)

For the sake of clarity only the result obtained for the second example (see Fig. 3) is presented in Fig. 4. In the case of the first example each of 116 fragments generates a quadruplet which is presented as a circle in the 2D image and the final figure is unreadable.

Because the number of spheres is usually too large, the set of quadruplets T ought to undergo the process of clustering. A simple clustering process [9] is used for set $T \subset P$ of quadruplets obtained for each fragments Q_i . Let us denote elements of T in the following way:

$$t_1 = (x_1, y_1, z_1, r_1), t_2 = (x_2, y_2, z_2, r_2), \dots, t_n = (x_n, y_n, z_n, r_n). \tag{15}$$

First assume that elements t_1, t_2, \dots, t_n form the initial cluster. We can calculate the co-ordinates of centre $c = (x_c, y_c, z_c, r_c)$ for a given cluster and the parameter Θ denoting the cluster concentration:

$$x_c = \frac{1}{n} \sum_{i=1}^n x_i, \quad y_c = \frac{1}{n} \sum_{i=1}^n y_i, \quad z_c = \frac{1}{n} \sum_{i=1}^n z_i, \quad r_c = \frac{1}{n} \sum_{i=1}^n r_i, \quad \Theta = \frac{1}{n} \sum_{i=1}^n \|c - t_i\|, \tag{16}$$

where $\|\cdot\|$ is the assumed norm within parameters space P .

If the calculated concentration Θ is greater than the assumed threshold value, Θ_{thresh} , the clustering algorithm is used. Otherwise, the initial cluster is treated as the final result of clustering.

CLUSTERING ALGORITHM – partitioning process of a given cluster $T = \{t_1, \dots, t_n\}$, where $(n > 1)$, is as follows:

Step 1: Choose two different initial centres of new clusters for a given cluster, e.g.:
 $c_1(1) = t_1$ and $c_2(1) = t_2$.

Step 2: With the j -th iteration step, divide set T into two subsets: $T_1(j)$ and $T_2(j)$ (transitory clusters) according to the following rule:

$$\begin{cases} t_i \in T_1(j) & \text{in the case of } \|t_i - c_1(j)\| < \|t_i - c_2(j)\|, \\ t_i \in T_2(j) & \text{otherwise} \end{cases}, \quad i = \overline{1..n}. \quad (17)$$

Step 3: On the basis of step 2, new cluster centres are determined as follows:

$$c_1(j+1) = \frac{1}{n_1} \sum_{t \in T_1(j)} t, \quad c_2(j+1) = \frac{1}{n_2} \sum_{t \in T_2(j)} t, \quad (18)$$

where n_1, n_2 are powers of sets T_1, T_2 , respectively.

Step 4: If $c_1(j+1) = c_1(j)$ and $c_2(j+1) = c_2(j)$ then END, else go to step 2.

If the application of the above algorithm does not cause the required concentration Θ determined in formula (16), the cluster undergoes the process of clustering again. Otherwise, two obtained clusters are treated as the final result.

It is necessary to verify every quadruplet $t_i = (x_i, y_i, z_i, r_i)$ obtained before and after the clustering. The verification should be performing within the following neighbourhoods:

$$x = x_i \pm \Delta, \quad y = y_i \pm \Delta, \quad z = z_i \pm \Delta, \quad r = r_i \pm \Delta, \quad \text{where } \Delta \in N. \quad (19)$$

This means that we must check every position of each sphere corresponding to each quadruplet on the real image. Before clustering we must look only for the best position of each sphere, as after clustering it is necessary to eliminate those quadruplets that do not correspond with spheres in the input image. The final result for the considered example is shown in Fig. 5 (compare Fig. 1).

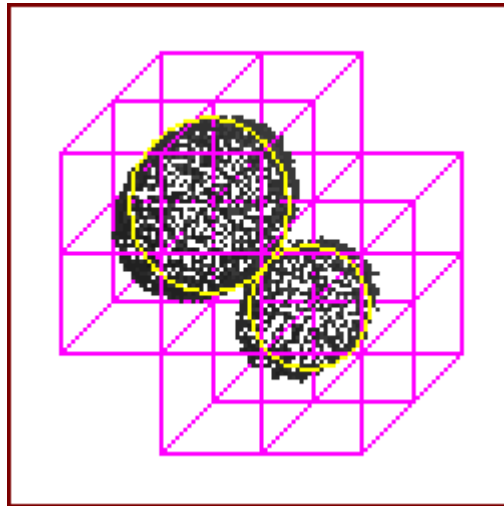


Fig. 5. The final result obtained for the considered example (see Fig. 4)

7 Conclusions and Further Research

The presented method is much faster than ordinary Hough transform and only 1D accumulator array is required. Nevertheless 3D images analysis is extremely computationally complex and takes minutes or even hours using standard PC. For comparison, analysis of the input image in Fig. 1 took less than a second.

It is necessary to observe that in the presented method the number of calculated Hough transforms corresponds to the number of fragments the image has been segmented into. As these are computationally independent the Hough transforms may be calculated in a parallel way.

The author is currently working on applying the presented technique in the case of straight lines and planes detection in 3D images. In principle, at this stage methods dedicated to 3D images have an academic thinking character.

References

- [1] Bandera A., Perez-Lorenzo J. M., Bandera J. P., Sandoval F.: *Mean shift based clustering of Hough domain for fast line segment detection*. PRL (27), No. 6, 2006, pp. 578-586.
- [2] Deans S. R.: *Hough transform from the Radon transform*. IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 3, no. 2, 1981, 185-188.
- [3] Hough P. V. C.: *Method and means for recognizing complex patterns*. U.S. Patent 3,069,654, Dec. 18, 1962.
- [4] Illingworth J., Kittler J.: *A survey of the Hough Transform*. Computer Vision, Graphics and Image Processing 44, 1988, pp. 87-116.
- [5] Leavers V. F.: *Shape Detection in Computer Vision Using the Hough Transform*. Springer, London 1992.
- [6] Li H., Lavin M. A., LeMaster R. J.: *Fast Hough transform: a hierarchical approach*. Computer Vision, Graphics, and Image Processing, vol. 36, 1986, 139-161.
- [7] Palmer P. L., Kittler J. V., Petrou M.: *An Optimizing Line Finder Using a Hough Transform Algorithm*. CVIU (67), No. 1, July 1997, pp. 1-23.
- [8] Radon J.: *Über die Bestimmung von Funktionen durch ihre Integralwerte langs gewisser Mannigfaltigkeiten*. Berichte Sachsische Akademie der Wissenschaften Leipzig, Math Phys Kl., 69, pp 262-267, 1917.
- [9] Tou J. T., Gonzalez R. C.: *Pattern recognition principles*. Addison-Wesley Publishing Company, London 1974.
- [10] Zorski W.: *Application of the Hough transform with a clustering technique to segmentation of digital images*. Machine Graphics & Vision, 5, 1996, pp. 111-121.
- [11] Zorski W.: *The Hough Transform Application Including Its Hardware Implementation*. Advanced Concepts for Intelligent Vision Systems: Proceedings of the 7th International Conference, Lecture Notes in Computer Science, Springer-Verlag Vol. 3708/2005, pp.460-467.
- [12] Zorski W., Foxon B., Blackledge J., Turner M.: *Irregular Pattern Recognition Using the Hough transform*. Machine Graphics & Vision, 9, 2000, pp. 609-632.
- [13] Zorski W., Foxon B., Blackledge J., Turner M.: *Application of the Circle Hough Transform with a Clustering Technique to Segmentation*. Image Processing II, Horwood Publishing, England 2000, pp. 339-348.

Context-Based Scene Recognition Using Bayesian Networks with Scale-Invariant Feature Transform

Seung-Bin Im and Sung-Bae Cho

Dept. of Computer Science, Yonsei University
134 Shinchon-dong, Sudaemoon-ku, Seoul 120-749, Korea
envymask@sclab.yonsei.ac.kr,
sbcho@cs.yonsei.ac.kr

Abstract. Scene understanding is an important problem in intelligent robotics. Since visual information is uncertain due to several reasons, we need a novel method that has robustness to the uncertainty. Bayesian probabilistic approach is robust to manage the uncertainty, and powerful to model high-level contexts like the relationship between places and objects. In this paper, we propose a context-based Bayesian method with SIFT for scene understanding. At first, image pre-processing extracts features from vision information and objects-existence information is extracted by SIFT that is rotation and scale invariant. This information is provided to Bayesian networks for robust inference in scene understanding. Experiments in complex real environments show that the proposed method is useful.

1 Introduction

Scene understanding is the highest-level operation in computer vision, and it is a very difficult and largely unsolved problem. For robust understanding, we must extract and infer meaningful information from image. Since a scene consists in several visual contexts, we have to recognize these contextual cues and understand their relationships. Therefore, it might be a good approach to start with extracting basic contexts like “where I am” or “what objects exist” in the scene for robust understanding. If we successfully extract these meaningful cues, we can provide them to higher level context understanding.

High-level context, like the correlations between places and objects or between activities and objects, is a key element to solve image understanding problem. For example, a beam-projector usually exists in a seminar room and a washing stand exists in a toilet. This contextual information helps to disambiguate the identity of the object and place despite the lack of sufficient information. Contextual scene recognition is based on common knowledge such as how scenes and objects are organized.

Visual information is powerful and crucial, whereas it is uncertain due to motion blur, irregular camera angle, bad lighting condition, etc. To overcome it, we need a sophisticated method that is robust to uncertainty. Bayesian network (BN) might be suitable for modeling in the domain of image understanding, since probabilistic

approach has the characteristic that is robust to inference in various directions and operable to uncertain data [1].

Probabilistic approach has attracted significant attention in the area of vision-based scene understanding. Torralba *et al.* proposed a method to recognize the place using hidden Markov model with global vectors collected from images and use them as context information to decide the detection priorities [2]. This approach is useful to make detection more efficient but the errors are inherited from the place recognition systems. Marengoni *et al.* tried to add the reasoning system to Ascender I which is the system to analyze aerial images for detecting buildings. They use hierarchical Bayesian networks and utility theory to select proper visual operator in the given context, and they could reduce computational complexity [3]. J. Luo, *et al.* proposed that Bayesian framework for image understanding [4]. In this approach, they used low-level features and high-level symbolic information for analyzing photo images.

In the meantime, there are many studies for solving object recognition problem. T. M. Strat and M. A. Fischler assumed that objects were defined by small number of shape models and local features [5]. D. G. Lowe proposed Scale-Invariant Feature Transform (SIFT) that extracts local feature vectors that are robust to image rotation and variation of scale [6]. SIFT shows good performance in extracting objects-existence but performance deteriorates if object has scanty texture element. Because performance of the object recognition algorithms is subject to low-level feature extraction results, we need a method that not only adopts low-level features but also uses high-level contexts.

In this paper, we propose a context based image understanding methodology based on Bayesian belief networks. The experiments in real university environment showed that our Bayesian approach using visual context based low level feature and high level object context which extracted by SIFT is effective.

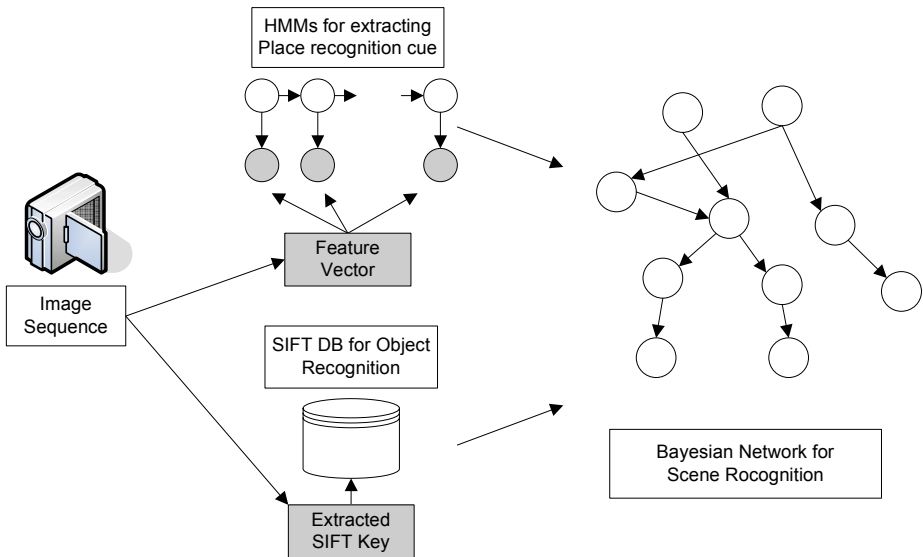


Fig. 1. An overview of Bayesian scene recognition

2 Context-Based Scene Recognition

In this section we describe the recognition of places and objects based on context. At first, we explain global feature extraction and HMMs learning, and describe object recognition with SIFT. Finally, context-based Bayesian network inference will be illustrated. The overview of the proposed method is shown in Fig 1.

2.1 Visual Context-Based Low-Level Feature Extraction

It would be better to use features that are related to functional constraints, which suggests to examine the textural properties of the image and their spatial layout [2]. To compute texture feature, a steerable pyramid is used with 6 orientations and 4 scales applied to the gray-scale image. The local representation of an image at time t is as follows:

$$v_t^L(x) = \{v_i, k(x)\}_{k=1, N}, \text{ where } N = 24 \tag{1}$$

It is desirable to capture global image properties, while keeping some spatial information. Therefore, we take the mean value of the magnitude of the local features averaged over large spatial regions:

$$m_t(x) = \sum_x |v_t^L(x')| w(x'-x), \text{ where } w(x) \text{ is the averaging window} \tag{2}$$

The resulting representation is down-sampled to have a spatial resolution of 4x4 pixels, leading to the size of m_t as 384(4 x 4 x 24), whose dimension is reduced by PCA (80 PCs).

Then, we have to compute the most likely location of the visual features acquired at time t . Let the place be denoted as $Q_t \in \{1, \dots, N_p\}$ where $N_p = 5$. Hidden Markov model (HMM) is used to get place probability as follows:

$$\begin{aligned} P(Q_t = q | v_{1:t}^G) &\propto p(v_{1:t}^G | Q_t = q) P(Q_t = q | v_{1:t-1}^G) \\ &= p(v_{1:t}^G | Q_t = q) \sum_{q'} A(q', q) P(Q_{t-1} = q' | v_{1:t-1}^G), \end{aligned} \tag{3}$$

where $A(q', q)$ is the topological transition matrix. The transition matrix is simply learned from labeled sequence data by counting the number of transitions from location i to location j .

We use a simple layered approach with HMM and Bayesian networks. This presents several advantages that are relevant to modeling high dimensional visual information: learning each level independently with less computation, and although environment changes, only first layer requires new learning with the remaining unchanged [7]. The HMM is for extracting place recognition and BNs are for high-level inference.

2.2 High-Level Context Extraction with SIFT

Scale-Invariant Feature Transform (SIFT) is used to compute high-level object existence information. Since visual information is uncertain, we need a method that has robustness to scale or camera angle change. It was shown that under a variety of reasonable assumptions the only possible scale-space kernel was the Gaussian function [6]. Therefore, the scale space of an image is defined as a function, $L(x, y, \sigma)$ that is produced by the convolution of a variable-scale Gaussian, $G(x, y, \sigma)$, with an input image, $I(x, y)$:

$$L(x, y, \sigma) = G(x, y, \sigma) * I(x, y), \quad (4)$$

where $*$ is the convolution operation in x and y , and

$$G(x, y, \sigma) = \frac{1}{2\pi\sigma^2} e^{-(x^2+y^2)/2\sigma^2} \quad (5)$$

To efficiently detect stable key-point locations in scale space, scale-space extrema in the difference-of-Gaussian function are convolved with the image, $D(x, y, \sigma)$, which can be computed from the difference of two nearby scales separated by a constant multiplicative factor k :

$$\begin{aligned} D(x, y, \sigma) &= (G(x, y, k\sigma) - G(x, y, \sigma)) * I(x, y) \\ &= L(x, y, k\sigma) - L(x, y, \sigma) \end{aligned} \quad (6)$$

Extracted key-points are examined in each scene image, and the algorithm decides that the object exists if match score is larger than a threshold.

In this paper, SIFT features of each object are extracted from a set of reference images and stored in an XML database. Each reference image is manually extracted from the training sequence set.

2.3 Context-Based Bayesian Network Inference

A Bayesian network is a graphical structure that allows us to represent and reason in an uncertain domain. The nodes in a Bayesian network represent a set of random variables from the domain. A set of directed arcs connect pairs of nodes, representing the direct dependencies between variables. Assuming discrete variables, the strength of the relationship between variables is quantified by conditional probability distributions associated with each node [8].

Consider a BN containing n nodes, Y_1 to Y_n , taken in that order. The joint probability for any desired assignment of values $\langle y_1, \dots, y_n \rangle$ to the tuple of network variables $\langle Y_1, \dots, Y_n \rangle$ can be computed by the following equation:

$$p(y_1, y_2, \dots, y_n) = \prod_i P(y_i \mid Parents(Y_i)) \quad (7)$$

where $Parents(Y_i)$ denotes the set of immediate predecessors of Y_i in the network.

BN used in this paper consists of 4 types of nodes: (1) ‘PCA Node’ for inserting global feature information of current place, (2) ‘Object Node’ representing object existence and correlation between object and place, and (3) ‘Current Place Node’ representing the probability of each place.

Let the place be denoted $Q_t \in \{1, \dots, N_p\}$ where $N_p = 5$, and object existence is denoted by $O_{t,i} \in \{1, \dots, N_{object}\}$ where $N_{object} = 14$. Place recognition can be computed by the following equation:

$$Current\ Place = \arg \max P(Q_t = q | v_{Lr}^G, O_{t,1}, \dots, O_{t,N_{object}}) \tag{8}$$

The BNs are manually constructed by expert, and nodes that have low dependency are not connected to reduce computational complexity. Fig. 2 shows a BN that is actually used in experiments.

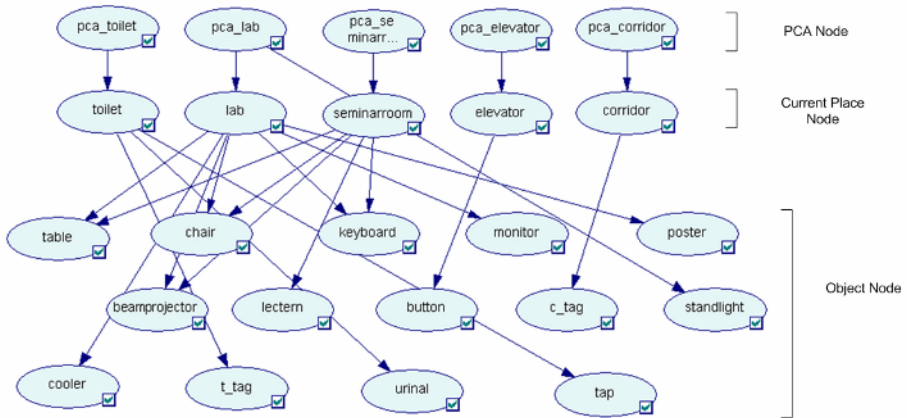


Fig. 2. A BN manually constructed for place and object recognition

3 Experimental Results

To collect input images, a USB mobile camera with notebook PC was used in the experiments. The camera was set to capture 4 images per second at a resolution of 320x240 pixels. The camera was set on a cap at the height of human sight, and the images were captured during user visits 5 different locations. The locations were visited in a fairly random order. We gathered 5 sequence data sets (one for training, others for testing) by the camera in the campus indoor environments. The sequences gathered contain many low quality images, due to motion blur, low-contrast and non-informative views, etc, but experimental results show that the proposed method overcomes these uncertainties.

Fig. 3 shows an experimental result that is the one of sequences that were used in our movements. The x-axis shows the flow of time and a solid line is the true places. Dots represent the probability of each inference result. The proposed method successfully

recognized the entire image sequences in general. However, during $t = 0$ to 100, in 'Elevator', the proposed method made several false recognitions, because of low-contrast and strong day light that passed through the nearby window. Due to scattered reflection, toilet and corridor also caused several false recognitions ($t = 320$ to 500).

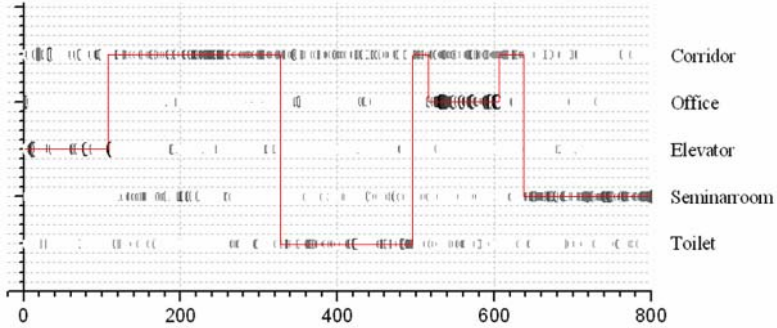


Fig. 3. One of the testing sequence result

Fig. 4 shows overall place recognition performance of the proposed method. The square dots show the place recognition results that used extracted low-level features only and diamond dots show the results of the method that used the BN with SIFT. It can be easily confirmed that the proposed method produces better performance. The hit rate of the proposed method increased 7.11% compared to the method that did not use BN. *Laboratory* shows highly increased recognition result since objects recognition performance by SIFT is good. On the other hand, *elevator* shows bad performance and smaller increase than other locations, because there is no particular object in elevator except elevator buttons, and bad light condition causes worse performance. In *toilet*, lack of the object existence information caused by diffused reflection made low recognition rate.

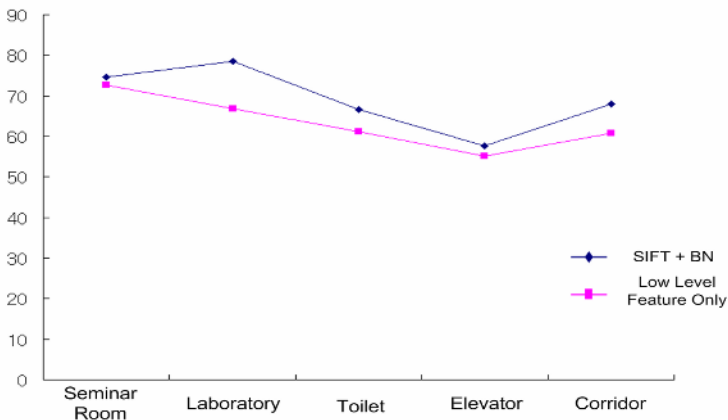


Fig. 4. Overall performance of each place recognition results

Fig. 5 shows the results of the SIFT object recognition. Objects with low texture features caused bad recognition results in the cases of *tap* and *urinal*. It can be easily confirmed that sufficient textual information makes good recognition result for the instances of the *keyboard* and *poster*. Fig. 6 shows the object recognition results of the proposed method. If the inferred objects-existence probability is larger than 75% or SIFT detects the object, the proposed method decides that object exists. Overall recognition score shows better results and recognition performance of objects that were not recognized by SIFT is increased especially (*monitor*, *urinal*). In addition, occluded objects were detected by Bayesian inference. However, it is a defect that false detection rate is increased in some objects.

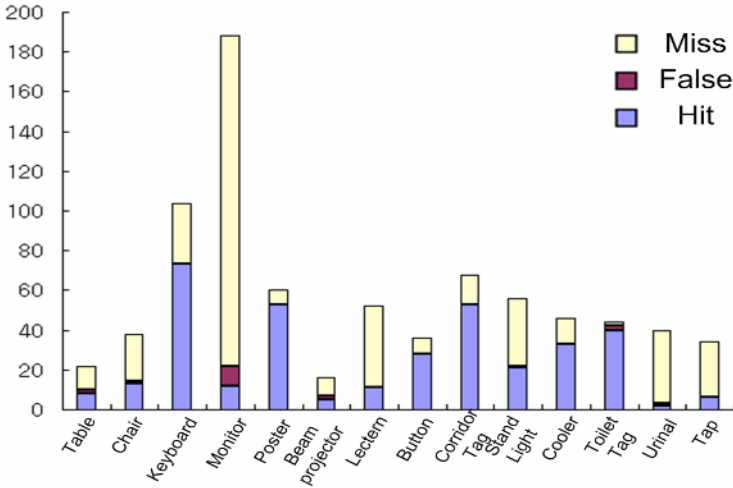


Fig. 5. Objects recognition results by SIFT

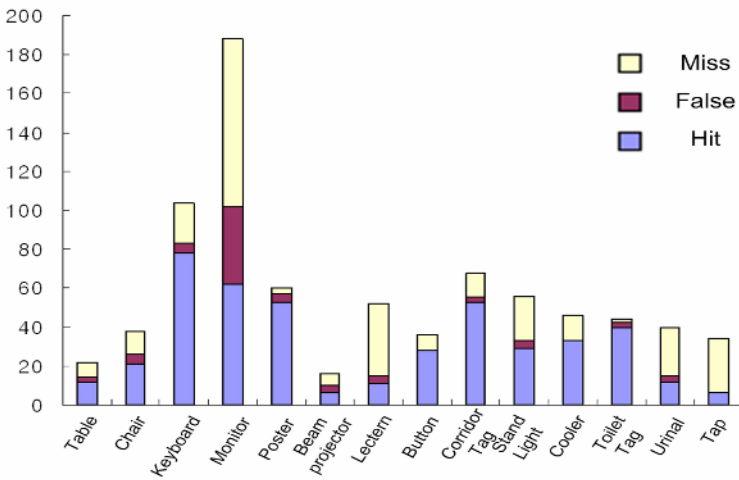


Fig. 6. Objects recognition results by the proposed method

4 Conclusions and Future Works

We have verified that the context-based Bayesian inference for scene recognition shows good performance in the complex real domains. Even though the global feature information extracted is the same, the proposed method could produce correct result using contextual information: relationship between object and place. But SIFT algorithm showed low performance when objects had insufficient textual features, and this lack of the information caused to the low performance of scene understanding. To overcome it, we need a method that disjoints objects with ontology concept, and extracts SIFT key-points in each component. Besides, we could easily adopt more robust object recognition algorithm to our method.

In the future works, we are under going to use the dynamic Bayesian network that represents previous state in scene understanding. Also, the application of the proposed method to real robot will be conducted.

Acknowledgments. This research was supported by the Ministry of Information and Communication, Korea under the Information Technology Research Center support program supervised by the Institute of Information Technology Assessment, IITA-2005-(C1090-0501-0019).

References

1. P. Korpipaa, M. Koskinen, J. Peltola, S. Mäkelä, and T. Seppänen "Bayesian approach to sensor-based context awareness," *Personal and Ubiquitous Computing Archive*, vol. 7, no. 4, pp. 113-124, 2003.
2. A. Torralba, K.P. Murphy, W. T. Freeman and M. A. Rubin, "Context-based vision system for place and object recognition," *IEEE Int. Conf. Computer Vision*, vol. 1, no. 1, pp. 273-280, 2003.
3. M. Marengoni, A. Hanson, S. Zilberstein and E. Riseman, "Decision making and uncertainty management in a 3D reconstruction system," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, no. 7, pp. 852-858, 2003.
4. J. Luo, A.E. Savakis, A. Singhal, "A Bayesian network-based framework for semantic image understanding", *Pattern Recognition*, vol. 38, no. 6, pp. 919-934, 2005.
5. T.M. Strat and M.A. Fischler, "Context-based vision: Recognizing objects using information from both 2-D and 3-D imagery," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 13, no. 10, pp. 1050-1065, 1991.
6. D.G. Lowe, "Distinctive image features from scale-invariant keypoints," *Intl. J. Computer Vision*, vol. 60, no. 2, pp. 91-110, 2004.
7. N. Oliver, A. Garg and E. Horvitz, "Layered representations for learning and inferring of-office activity from multiple sensory channels," *Computer Vision and Image Understanding*, vol. 96, no. 2, pp. 163-180, 2004.
8. R.E. Neapolitan, *Learning Bayesian Network*, Prentice hall series in Artificial Intelligence, 2003.
9. J. Portilla, and E.P. Simoncelli, "A parametric texture model based on joint statistics of complex wavelets coefficients," *Intl. J Computer Vision*, vol. 40, no. 1, pp. 49-71, 2000.
10. G.F. Cooper and E. Herskovits, "A Bayesian method for the induction of probabilistic networks from data," *Machine Learning*, vol. 9, no. 4, pp. 309-347, 1992.

A Portable and Low-Cost E-Learning Video Capture System

Richard Y.D. Xu

School of Information Technology, Charles Sturt University
Bathurst NSW 2795 Australia
rxu@csu.edu.au

Abstract. In the recent times, many computer vision supported e-learning applications have been constructed, to provide the participants with the automated and real-time camera control capabilities. In this paper, we describe a portable and single-PC based instructional video capture system, which incorporates a variety of computer vision techniques for its video directing and close-up region specification. We describe the technologies used, including the laser-pointer detections, instructor's lip tracking and individual teaching object recognition. As the same time, we also explain how we have achieved both *low-cost* and *portability* property in our design.

1 Introduction

The recent advancements in computer vision technologies and video-capturing devices (both static and computer-controlled motorized Pan-Tilt-Zoom cameras) have led to the increase in the number of automated e-learning video capturing systems.

Some examples of the recent works in this area includes [1], which is a three cameras system and its control rules are based on a set of estimations to teacher's position, face direction, hand position and amount of characters on the blackboard. In [2], the authors have used hand tracking techniques to automate the camera operations, where its rules are based on how user presents objects picked up from the desktop. The system in [3] is comprised of a static and tracking camera to capture the slides and the speaker, and direct video-shooting based on changes occurred on the stage, the screen, the lectern and the presenter positions. We have also listed a detailed survey on the subject [4] for the interested readers.

1.1 Motivation of Portability and Low Cost

Most current automatic video capture systems require multiple industrial camera [1-3] and they are usually multiple PC-based. These systems clearly do not meet the requirements for many low-cost e-learning scenarios. Our aim is to design an inexpensive camera system, which requires only single PC processing, and uses consumer-type camera.

Secondly, the cameras in the existing systems [1-3] are installed to a classroom, typically mounted to the ceiling. This design allows the position of the observing camera to remain unchanged, which is necessary for the stereo-vision tracking methods based on pre-calibrated camera parameters. Our second motivation is to design a ready-to-use e-learning video capture system, which can be ported easily from one instructor room to another.

Both the *low-cost* and *high portability* property are important in many e-learning scenarios [4]. To achieve these two purposes, our research, therefore aims to apply a set of alternative computer vision techniques, to compensate the lack of stereo-vision, multiple PC and industrial camera.

1.2 Motivation on Versatility

Most current-day vision-based camera control is driven by the instructor's actions and changes occurred in whiteboard writings. Our research also considers that the interaction should also extend to other modalities, including the teaching equipments; instructor's gestures and laser pointer interaction.

1.3 Hardware

In our system, the PTZ and static camera are placed sufficiently close at eye level shown in Fig 1.



Fig. 1. PTZ and static camera configurations

The PTZ camera system consists of an Eagletron PowerPod base used in conjunction with a household camera, which cost less than USD\$599, running on standard PC (Pentium M, 1.7 GHZ, 512M RAM). This setup satisfies both our low-cost and portability goal. However the hardware's mechanical imprecision is a disadvantage, for which we have designed a unique mechanical convergence algorithm for its movement control, illustrated in Section 4.

The rest of this paper is organized as follows, in section 2, we illustrate our methods to obtain camera close-up region, using both laser pointer guidance and hand-held individual object recognition. In section 3, we will present our lip tracking result which supports for the audiovisual speech command. In section 4, we will illustrate our PTZ camera movement control.

2 Instructor Specified Close-Up Region

In an e-learning session, camera needs to zoom into several subjects, so that the remote participants can have a look-up view. The subjects include *the instructor*, *the teaching object* and *a static region* (for example, the whiteboard). While person (instructor) close-up tracking has been addressed in many literatures, in this Section, we will only discuss the *teaching object* and a *static region* specification.

2.1 Hand-Held Teaching Object Region Specification

For teaching object that is small in size, we have introduced a set of natural camera control and region definition, called *individual object interaction* techniques [5]. The correct close-up region is defined after a “*recognized*” teaching object is presented by the instructor.

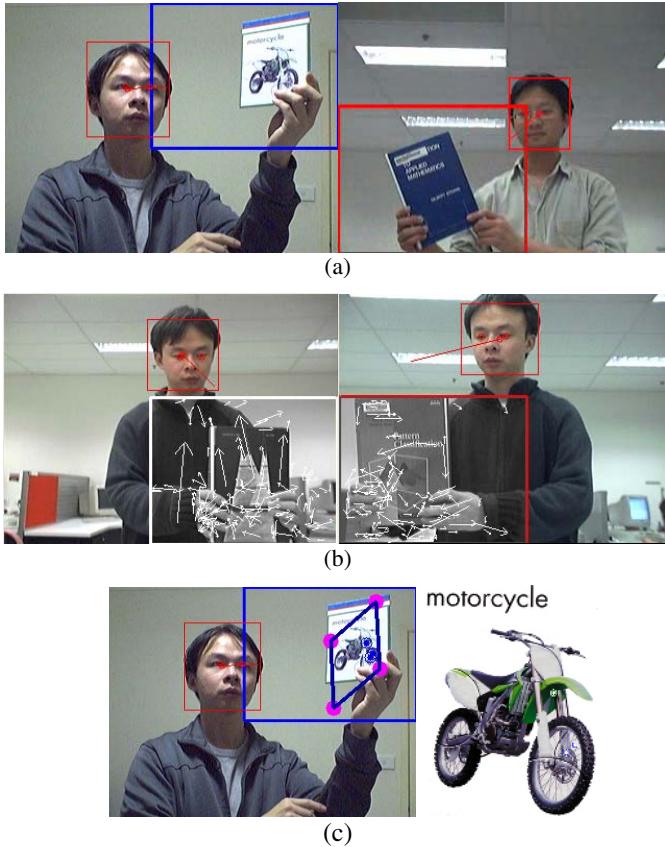


Fig. 2. Individual object interaction (a) ROI candidates determined from facial pose. (b) The SIFT feature is computed within the ROI (c) The matched object from the image database.

Abstractly, the process is achieved by first estimate region of interest (ROI) from presenter's head pose, Shown in Fig 2.a, the rectangle region is determined from the instructor's pose.

We then compute within the ROI, the Scale Invariant Feature Transform (SIFT) [6] matching, shown in Fig 2.b, where the arrows indicate the SIFT features' position, scale and orientation.

We finally compare the SIFT feature generated in ROI against the SIFT feature stored in the image database, using the Approximated Nearest Neighbor (ANN) method and determines the teaching object that instructor is currently holding, Fig 2.c. The details of this work is given in [5].

2.2 Static Region Specification Using Laser Pointer

Similar to many laser pointer interaction [7, 8], in our system, an instructor can specify the zoom-in region by drawing virtual ellipse around the ROI, show in Fig 3.

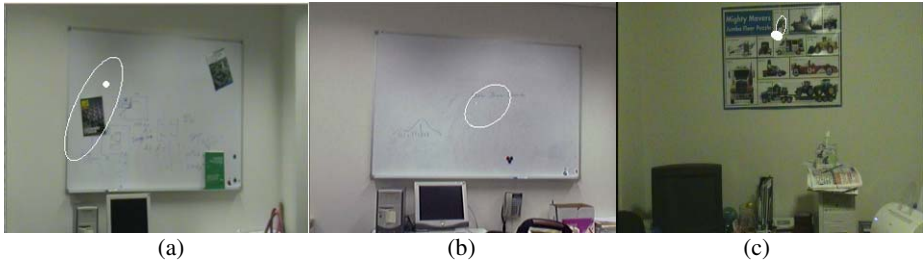


Fig. 3. Instructor-specified virtual ellipse around a. teaching object, b. whiteboard region and c. poster region

However, robust laser pointer detection is not as trivial as “*finding a small red patch in the image*”. Apart from the background noises, such as reflection of metallic surfaces, the most difficult challenge results from a so-called CCD clipping phenomenon [7], where the laser pointer illumination on white surface appears on the camera CCD as a bright white patch.

Most cameras' CCD sensor is only capable capturing light with an intensity threshold, and it treats a white surface (for example, a whiteboard) close enough to that threshold. Therefore, the amount of intensity increase between the laser pointer illuminated patch and its “white” surroundings is not as significant compared with when laser pointer is projected onto a “darker” surface. For these reasons, previous laser pointer detection approaches employs manual preconfigured camera settings with simple RGB thresholding [7, 9] are extremely sensitive to light changes and has very low detection rates, which is replaced by machine-learning approach [10, 11].

We have used a spatiotemporal training strategy, where a bi-modal Gaussian Mixture Model (GMM) is used for both spatial and temporal features:

2.2.1 Temporal Features

The temporal training strategy is similar to [11], we have recorded the maximum absolute pixel differences between two consecutive video frames. Therefore, unlike training method used in [12], our method is insensitive to camera changes during real-time detection. This is inline with our portability goal.

In order to obtain the features, the instructor is required to project laser pointers over a training region or the entire scene. The instructor roughly requires covering half of the region with laser pointers, leaving enough examples in the region both with and without illumination.

Instead of recording single pixel difference used in [11], where singly detected pixel are grouped together later to form laser pointer region, in our method, we have simplified the process by using integral image features, where average intensity of a small region (3 x 3) pixels is calculated in constant time. This training process takes about 10 to 15 seconds depending on the size of training regions used. We then obtain a histogram calculated using both R and B components. This is shown in Fig 4.a. In this histogram, we can clearly see high values, which resulted from regions being illuminated with laser pointer some time during training. The lower values correspond to regions without illumination at all. Due to camera’s CCD noise, these values are slightly above 0, generally around 10 depends on how “fluctuate” the CCD camera is.

From this histogram, a clear two modes are identified. Intuitively, a bi-modal Gaussian Mixture Model (GMM), a commonly used pattern recognition method [13], can sufficiently model such histogram. The equation for GMM is:

$$GMM(\mathbf{x}) = \sum_{k=1}^K w_k \cdot g_{(\mu_k, \Sigma_k)}(\mathbf{x}) \tag{1}$$

where g is the Gaussian distribution

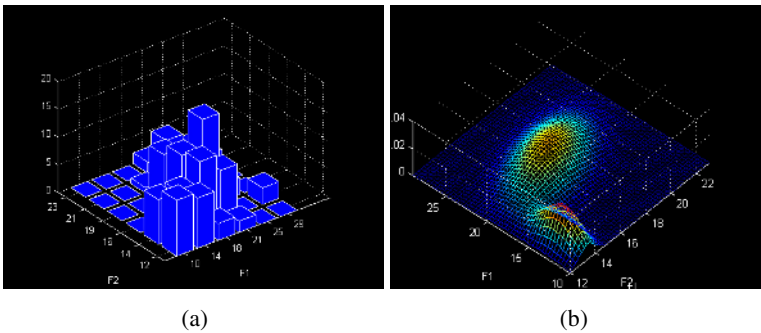


Fig. 4. Modeling temporal features: a. is the Histogram; b. is the corresponding bi-modal GMM fittings obtained

Since we can also guess its initial values of the two modes, being close to 10 and 30 respectively, a GMM model can be fitted accurately using Expectation-Maximization (EM) algorithm. The result of GMM fitting is shown in Fig 4.b. The weight of each Gaussian w_k is not in our interest.

A threshold value is then drawn from this GMM model which maximally discriminate the correctly classified samples from noise ones during real-time detection.

The results achieved from the temporal feature training is promising, during detection, only one or two outliers are noticed over time from the whiteboard regions, mostly due to light reflections on the metallic whiteboard.

2.2.2 Spatial Features

The occasional outliers from the temporal classification can be removed easily by heuristic methods, such as using location threshold [11]. However, for our application, an outlier-less detection is more desirable. To further remove any outliers caused by the occasional noises, we combined the result with the spatial features training.

To obtain the spatial features, we have used integral images [10] feature. Integral images feature are commonly used, such as in Adaboost training of facial image [14]. This method computes much faster than convolution filters, where the image sum of any size region can be obtained from three addition/subtraction operation [14].

In our work, we have used spatial features from the R component alone. The feature sum is calculated a from small region I_{small} subtract from its surrounding I_{large} . I_{small} is typically chosen to be 3x3, and I_{large} is typically chosen to be 6x6.

Like the temporal case, in order to collect both examples in regions with and without laser pointer illuminations, we have recorded the maximum feature value in the training region over time. The training result is collected at the same time of the temporal features. A strong bi-model histogram can be once again obtained. In regions where no laser pointer has been illuminated before, $I_{small} - I_{large}$ is closer to 0 with some noise obtained. The other mode represents the regions where laser pointer is illuminated some point during training. The GMM fittings were once again used to discriminate between the two modes.

Spatial features generate more outliers than that of the temporal one. However, during real-time detection, we have used spatial features as an outlier remover after temporal features detection. In other word, spatial features are only used, checking if detected location using temporal features is valid. For this reason, spatial features do not involve computation over the entire video frame.

2.2.3 Laser Pointer Detection Result

2.2.3.1 Robustness. By using our combined spatial and temporal training, we have achieved an outlier-free detection in 98% of the testing cases. We have noted that there are a few un-detections occur from time to time. This is due to the strict constraints we have placed upon our detection algorithm, since false detection is more prohibitive than un-detection in our application. However, false detection is very rare.

2.2.3.2 Efficiencies. Unlike works in [11], where much post-filtering is required. Our laser pointer detection is self-sufficient, and most of calculations are achieved using integral images features. The efficiency of our laser pointer detection algorithm is shown in Fig 5, where the execution time is around 26 ms and is almost constant across times. This allows us to achieve real-time detection.

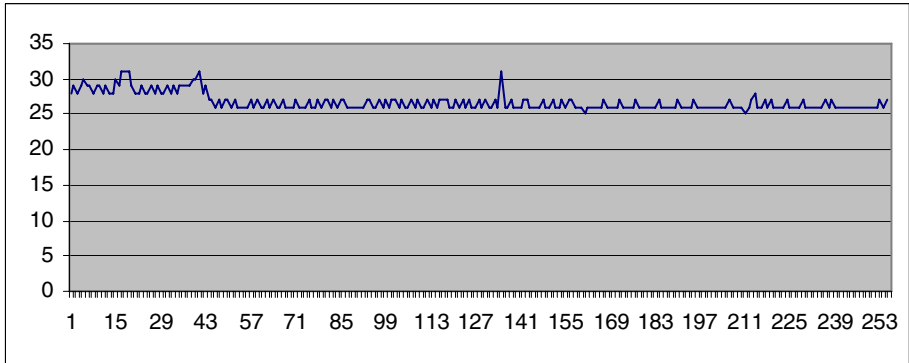


Fig. 5. Execution time for laser pointer detection

3 Lip Tracking to Support Instructor's Voice Recognition

For the instructor-directed camera controls, we have formulated a multimodal approach. The first one is to use hand-gesture and the second one is to use voice command recognition, to issue simple commands, such as, “*zoom laser*”, “*track instructor*” corresponds to “*zoom according to laser pointer*”, “*follow the instructor*” respectively.

In order to achieve more robust voice recognition under noisy environment, we are experimenting combined audiovisual voice recognition with additional lip tracking information. In our current work, we have used Support Vector Machine (SVM) to detect lip region, fine-tuned from Haar-like feature detected mouth region.

3.1 Mouth Region Detection

We obtain the close-up facial image from the PTZ camera view, using PTZ camera convergence algorithm described in Section 4. We have used a standard Viola and

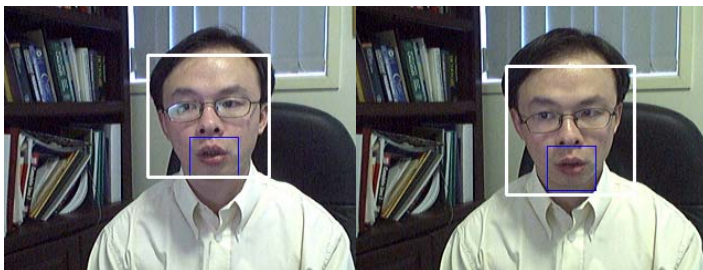


Fig. 6. Mouth region detection using Haar-like feature detection, white rectangle is detected frontal face, notice the rectangles fluctuate even when the person is stationary

Jones [14] to detect the mouth region. The detection is executed in two steps. A frontal face training library is used for detecting the face initially, and then followed by mouth detection using the mouth training library, shown in Fig 6.

We notice that although the detected region contains the mouth, its location is not always at the centre of the detected region. In addition, detected mouth area fluctuates even when the instructor is stationary. From the detected mouth region, we then try to fine-tune the lips area using a Support Vector Machine (SVM) classifier.

3.2 The Lip Tracking

The features we use for SVM classification are based on both colour and shape. When only three-dimensional colour features (in YCrCb colour space) are used, the results often contain high percentage of noises. This is indicated in Fig 7.a, in a Haar-like feature detected mouth region, part of a nose and shadows are classified as mouth region, even they are labelled $y = -1$ during training.

When the additional shape features are introduced in the training set, the classification generated much less noises than the ones shown in Fig 7.a. The improved result is shown in Fig 7.b:

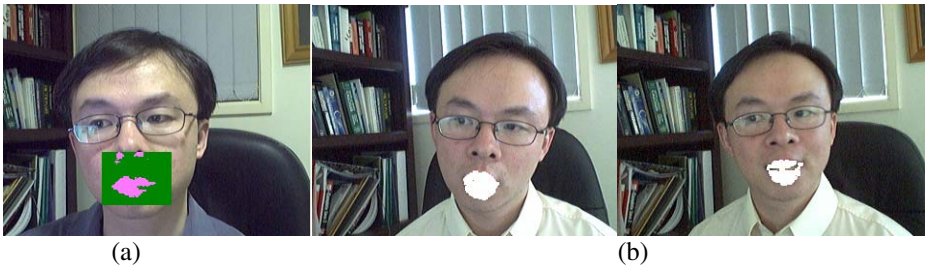


Fig. 7. (a) SVM classifier using three dimensional colour feature training, (b) Lip region results after repeated SVM classification with refined



Fig. 8. The lip tracking result, http://www-staff.it.uts.edu.au/~richardx/lip_richard.avi

After classification, the final mouth shape is determined from active contour method. The result is shown in Fig 8. We have also experimented with several other

SVM kernels tested with eight different persons. The Radial Basis Function (RBF) has achieved most robust result.

4 Camera Movement

As stated in the Introduction, for the portability reason, we are restricted from using 3D stereo triangulation for person tracking. Therefore, camera zoom-in operation is achieved through our own mechanical convergence algorithm.

4.1 Estimating PTZ Camera Direction from Static Camera View

We control the PTZ camera's pan and tilt movement initially from static camera information. As a consequence, when the tracking subject becomes more stationary, PTZ camera's mechanical convergence starts from a direction closer to the subject. To achieve depth calculation using monocular (static camera), we have used a similar method to [15], where we measure the detected face sizes in the static camera across different depths. During real-time, detected face sizes serve as depth approximation.

This method has low precision, but it achieves our purpose, where in most of the times, the instructor's face is included in the PTZ camera view even though the face may not be at the centre. It then allows PTZ camera's mechanical convergence procedure to begin when static camera detects instructor's movements is slowed.

In Fig 9, we have shown the tracking result based on this technique:



Fig. 9. PTZ camera's initial direction control by static camera view

The first five screen captures shows PTZ camera's movement controlled by detected face sizes in static camera view. During this time, the person has large movement and PTZ camera is not exactly pointing into the direction of instructor's face, but contains it within its camera view. In the last screen capture, instructor remains stationary for a five-second period to allow PTZ camera complete its

mechanical convergence. More camera results is found http://www-staff.it.uts.edu.au/~richardx/PTZ_camera.WMV. The tracking after initial face detection is based on mean-shift based color tracking [16].

4.2 PTZ Camera's Mechanical Convergence

Before the zoom-in operation takes place, the PTZ camera needs to pan and tilt until the instructor's face is in its centre. Our contribution lies in our semi-passive, mechanical convergence control algorithm. This algorithm allows the low-precision PTZ camera base to perform “centering” operation using the face as a reference object. By accurately track the selected region while continuously updating the position of PTZ camera's current centre view, its mechanical movements can effectively be converged, such that the selected region and the centre of camera view will overlap

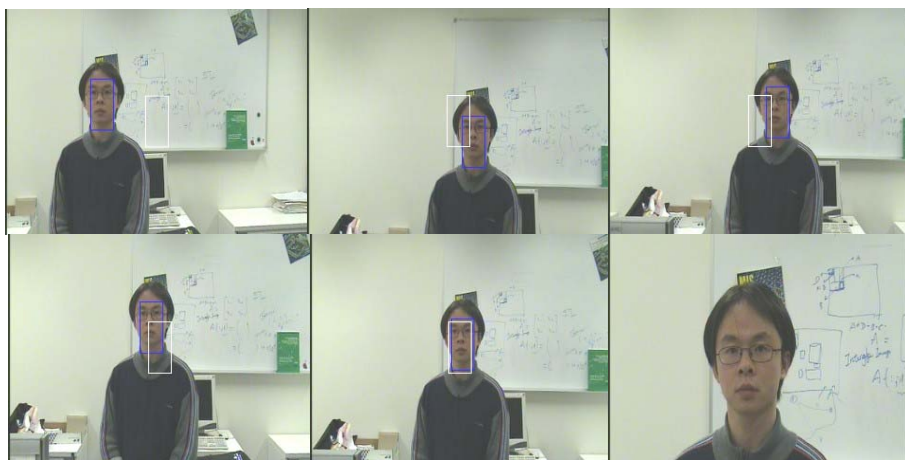


Fig. 10. Result of PTZ camera's mechanical convergence, prior to optical zoom-in operation

The algorithm begins with larger movements. When it is getting close to an object or has ‘passed’ over the object, its motion becomes slower and in a reverse direction. In Fig 10, we have shown the result of the PTZ camera view using the above algorithm. Notice that the initial movement is larger, then it becomes slower and in a reverse direction when the object (in blue) “passed over” the centre view (in white). More of the tracking results can be found on:

<http://www-staff.it.uts.edu.au/~richardx/PTZCameraZoom.WMV>

5 Conclusion

In this paper, we have described a set of monocular computer vision methods, to which we have use to detect the camera close up region; camera movement control and support more robust instructor audiovisual speech recognition. The techniques we have applied allow the system to be portable, and all the computer vision processing is achieved using single PC, which also compensates for the low-cost, low-accuracy

camera system. Our future work is to further extend the interaction methods, such as more robust instructor-led body gesture, and also combining cinematography rules into instructional video presentation.

References

- [1] A. Shimada, A. Suganuma, and R. Taniguchi, "Automatic Camera Control System for a Distant Lecture Based on Estimation of Teacher's Behavior," International Conference on Computers and Advanced Technology in Education, 2004.
- [2] M. Ozeki, Y. Nakamura, and Y. Ohta, "Automated camerawork for capturing desktop presentations - camerawork design and evaluation in virtual and real scenes," 1st European Conference on Visual Media Production (CVMP), 2004.
- [3] M. Bianchi, "Automatic video production of lectures using an intelligent and aware environment," 3rd international conference on Mobile and ubiquitous multimedia, College Park, Maryland, 2004.
- [4] R. Y. D. Xu and J. S. Jin, "Adapting Computer Vision Algorithms to Real-time Peer-to-Peer E-learning: Review, Issues and Solutions," To Appear in World Conference on E-Learning in Corporate, Government, Healthcare, and Higher Education, Vancouver, Canada, 2005.
- [5] R. Y. D. Xu and J. S. Jin, "Individual Object Interaction for Camera Control and Multimedia Synchronization," (to appear) 31st IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP 2006), Toulouse, France, 2006.
- [6] D. Lowe, "Distinctive image features from scale invariant key points," *International Journal of Computer Vision*, vol. 60, pp. 91-110, 2004.
- [7] D. R. Olsen and T. Nielsen, "Laser pointer interaction," Proceedings of the SIGCHI conference on Human factors in computing systems, Seattle, Washington, United States, 2001.
- [8] R. R. Eckert and J. A. Moore, "The classroom of the 21st century: The interactive learning wall," vol. 32, pp. 33-40, 2000.
- [9] J. Oh and W. Stuerzlinger, "Laser pointers as collaborative pointing devices," Graphics Interface 2002, 2002.
- [10] D. Olsen, "A design tool for camera-based interaction," Proceedings of the SIGCHI conference on Human factors in computing systems (CHI2003), 2003.
- [11] Y. Shi, W. Xie, G. Xu, R. Shi, E. Chen, Y. Mao, and F. Liu, "The Smart Classroom: Merging Technologies for Seamless Tele-Education," *Pervasive Computing*, vol. 2, pp. 47-55, 2003.
- [12] B. A. Ahlborn, D. C. Thompson, O. Kreylos, B. Hamann, and O. G. Staadt, "A practical system for laser pointer interaction on large displays," ACM Symposium on Virtual Reality Software and Technology (VRST 2005), 2005.
- [13] P. E. Hart and D. G. Stork, *Pattern Classification*: Wiley, 2001.
- [14] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," CVPR 2001, 2001.
- [15] K. Cheng and M. Takatsuka, "Real-time Monocular Tracking of View Frustum for Large Screen Human-Computer Interaction," Twenty-Eighth Australasian Computer Science Conference (ACSC2005), Newcastle, Australia, 2005.
- [16] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-Based Object Tracking," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 25, pp. 564-575, 2003.

On Building Omnidirectional Image Signatures Using Haar Invariant Features: Application to the Localization of Robots

Cyril Charron, Ouiddad Labbani-Igbida, and El Mustapha Mouaddib

Centre de Robotique, Electrotechnique et Automatique,
Université de Picardie Jules Verne, 7 rue du Moulin Neuf,
F-80039 Amiens Cedex, France
{cyril.charron, ouiddad.labbani, mouaddib}@u-picardie.fr

Abstract. In this paper, we present a method for producing omnidirectional image signatures that are purposed to localize a mobile robot in an office environment. To solve the problem of perceptual aliasing common to the image based recognition approaches, we choose to build signatures that greatly vary between rooms and slowly vary inside a given room. To do so, an invariant approach has been developed, based on Haar invariant integrals. It takes into account the movements the robot can do in a room and the omni image transformations thus produced. A comparison with existing methods is presented using the Fisher criterion. Our method appears to get significantly better results for place recognition and robot localization, reducing in a positive way the perceptual aliasing.

1 Introduction

The general localization problem is the most crucial perceptual problem in robotics. In several successful methods (like odometric map matching methods [3] or landmark based methods [5]) to robots localization, often called position estimation, the robot is given a (geometrical and metrical) map of the environment and its goal is to determine its position relative to this map given perceptions of the environment and the robot movements. However, it is not always possible or convenient to provide a geometrical map of the environment. Image-based localization approaches [12,20,27,7,16,9,15] prevent you from having to use a map and give a rough estimation of the robot's location by matching a set of views taken by the robot to reference views stored in previous experiments. These approaches, called image-based (or appearance-based) localization, are worthwhile to build topological maps and could be applied in exploration and video surveillance of a priori unknown environments.

In this paper, we are interested in computing omnidirectional image signatures that will help a robot to recognize images with the aim of locating itself in unknown and explored environments. The robot position estimation problem consists in finding the best match for the current image among the reference images. This can be a tricky problem if the environment displays symmetrical

structures like doors and corridors, so the current view will match not only the referred location image, but also all similar images giving *perceptual aliasing*.

The approach we propose in this paper deals with this problem, giving an elegant way to compute image signatures that enhance the discriminating power of the localization algorithm. The paper is organized as follows: We start with a review of the related work and, in Section 3, continue with the principles of building omnidirectional image signatures using Haar integral invariant features. In Section 4, we describe some results of the experiments in several indoor environments and compare them in Section 5 to related approaches we implemented. Finally, in Section 6 we conclude with a summary and an outline of ongoing work.

1.1 Related Works

Several research groups [12,20,27,7,16,9,15] have presented successful methods for image matching that differ mostly in the way the models are built to reduce the image vector by extracting relevant features. Most of them benefit from matching techniques (Fourier harmonics, eigenspace or invariant features) developed in the field of image retrieval. Actually, image retrieval systems aim to find images that are similar in appearance to an input query, from a large-size database. However, while a few bad matches are not a problem in image retrieval, a single bad match could lead the robot localization system to get lost and must therefore be strictly avoided for the localization task.

Image retrieval systems usually rely on histograms for the matching process. This is due to their compact representation of the images, their invariance to rotation (which is very interesting for omnidirectional images) and their very low sensitivity to small translations. Ulrich et al. [27] use histograms of omnidirectional images associated with a nearest-neighbor learning by unanimous voting, to localize a robot in a topological map with good confidence ratios. Color histograms have also been applied by [9] to localize a Rover robot in natural environments using omnidirectional images. But the histograms are not invariant when important movements are involved.

Other works, as [12,13], have issued image compact representations using subspaces of Fourier harmonics, i.e. they calculate the Fourier coefficients to represent images in a lower-dimensional subspace. These representations lack robustness since the Fourier transform is inherently a non-robust transformation. An occlusion in the image influences the frequency spectra in a non-predictable way and therefore arbitrarily changes the coefficients. Recently, Menegatti et al. [21] have applied global Fourier transform to extract the coefficients of the low frequency components of omnidirectional images grabbed by the robot. To overcome the lack of robustness in the case of perceptual aliasing, they used in [22] a Monte-Carlo localization technique and their system was able to track the position of the robot while it was moving and estimate its position without any prior knowledge on the real position.

In the case of the eigenspace approaches, you need to build a database model by computing the eigenvectors or the principal components [15,20,1,16,7]. The main interest here is to find an invariant representation to the omnidirectional

image rotations (taken at the same position under different orientations of the robot). The images are wrapped to cylindrical panoramic representations in that way a rotation of the original image is equivalent to a shift of the image plane deployed from the cylindrical image. Aihara et al. [1] use row-autocorrelated transforms of cylindrical panoramic images for indoors and outdoors localizations. The approach suffers from less accurate results for images acquired on novel positions, since by autocorrelating the images some of the information is lost. Moreover, any occlusion in the image may result in an erroneous localization. Pajdla and Hlaváč [23] propose to estimate a reference orientation from images with the Zero Phase Representation (ZPR). ZPR, in contrast to autocorrelation, tends to preserve the original image content while at the same time achieving rotational independence, as it orients images by zeroing the phase of the first harmonic of the Fourier transform of the image. The experiments indicate that images taken at nearby positions tend to have the same reference orientation. The method is however sensitive to variations in the scene and occlusions, since it only operates with one single frequency using a global transform.

The idea of invariant methods is to build image features that should exhibit invariance against different transformations on the scenes. SIFT¹ features, developed by Lowe [18,19], are invariant to image translation, scaling and rotation. They are also partially invariant to illumination changes. Variants of SIFT (Modified SIFT [2], Iterative SIFT [26]) have been proposed to reduce the computational efforts of the feature extraction and matching process, and applied in almost real time robot localization.

An interesting approach to *formally* define invariant signatures of images that undergo group transformations is *Haar integrals*. Firstly introduced by Schulz-Mirbach in [24], this invariant has been used, in case of euclidian motions, for image retrieval by Siggelkow [25], Halawani and Burkhardt [10,11], and for mobile robot localization by Wolf [14] (although the Haar integral was not explicitly used). The Haar integral invariant features could be extracted directly from raw images, without need to preprocessing such as segmentation or edge extraction.

In this paper, we apply Haar integral invariants to compute signatures on omnidirectional images and compare our method to representative works of the literature: Using Fourier transform [22] and Zero-phase representation [23]. The different methods are implemented in our localization system and tested in different indoor environments.

2 Omni-Appearance Based Localization Algorithm

The robot localization approach is a twofold procedure: In the setup stage, the robot takes a set of omnidirectional images at reference locations which form a good depiction of the environment (by following some training strategy or under the supervision of a human operator). Then, Haar invariant signatures to rotations and local translations in the scene are computed, which allows us

¹ Scale Invariant Feature Transform.

to represent every image of the training set with only low-dimensional invariant distributions.

At the running stage, while it moves in the environment, the robot acquires new images, computes their Haar signatures and then searches for the nearest signatures in the model built at the training set.

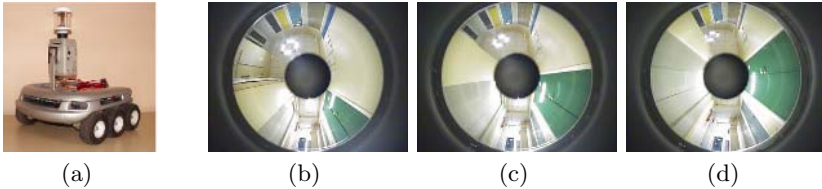


Fig. 1. **a:** The Koala robot and its omnidirectional sensor. **b,c** and **d** show images acquired by the robot's sensor during a translation.

The localization procedure is repeated for the different models compared of signature extraction: Fourier transforms [21], Zero phase representation [23] and our approach of Haar integral invariant distributions, in the same conditions. Figure 1 sketches an example of an indoor environment seen by the omnidirectional sensor of the robot.

3 Haar Invariant Signatures Extraction

3.1 The General Idea

The approach we develop is inspired by Haar integral invariant features introduced by Schulz-Mirbach in [24] and applied by Siggelkow [25] in image retrieval. The main difference lies in the nature of the transformations considered. Siggelkow deals with euclidian transformations of the images, given cyclic boundary conditions. These hypotheses on image transformations do not hold anymore when dealing with the geometry of omnidirectional sensors.

Haar integral is expressed as

$$I_{Haar}(\mathbf{x}) = \frac{1}{|G|} \int_G f[g(\mathbf{x})] dg \text{ with } |G| = \int_G dg \quad (1)$$

where G is the transformation group, and $g(x)$ the action of g , an element of G , on vector \mathbf{x} . It could be viewed as a course through the space of the transformation group parameters. In case of $\mathbf{x} = \mathbf{M}$ being an image, equation 1 suggests that the integral invariant feature is computed by first 1) applying kernel function f to each pixel in transformed image $g(\mathbf{M})$ then 2) summing up over all transformations of G and 3) normalizing the result to get a single representation of the invariant feature.

We generalize Haar integral features to integrate the complex transformations induced by the geometry of the sensor and transformed under the robot movements (translations and rotations) in the scene. We define distributions based on a partition of the constructed Haar integral features and build up histograms using these distributions.

3.2 The Camera Transformation Model

The robot is endowed with an omnidirectional sensor generating complex projective transformations. Omni-images are often projected back onto a cylinder, and then mapped back into a plane by an isometry so as to have them looking more similar to classical perspective images (Fig. 2(b)). This is the case in the work of Menegatti [21] or Pajdla [23].

Here we use the equivalence sphere model given by Geyer and Daniilidis [8]. They prove that central catadioptric projection can be modeled with the projection of the sphere to a horizontal plane from a point on the vertical axis of the sphere. Once the sensor has been calibrated, the raw image is projected onto this sphere, equivalent to the actual mirror from the point of view of the image formation process (Fig. 2(a)). Spherical image $\mathbf{M}_S(\theta, \varphi)$ (equation 2) has a topology which looks more adapted to the sensor properties than the raw image.

\mathbf{M}_S is formed by regularly meshing the sphere in (θ, φ) and interpolating at the corresponding points in the original image, $\mathbf{M}(u, v)$ using the projective equation (2):

$$\begin{cases} u = \cot(\theta/2) \cdot \cos(\varphi) \\ v = \cot(\theta/2) \cdot \sin(\varphi) \end{cases} \quad (2)$$

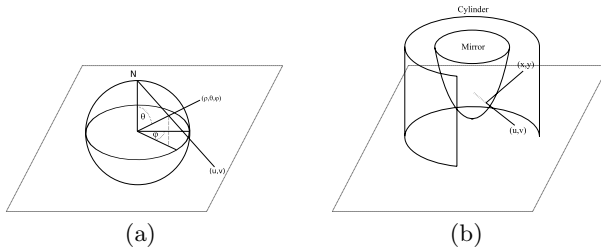


Fig. 2. a) The equivalent sphere. To 3D point (ρ, θ, φ) corresponds a point (u, v) in the raw image. b) The projection of the raw image onto a cylinder used by Menegatti and Pajdla.

3.3 The Transformation Group

Without loss of generality, the reference frame of the robot and that of the mirror can be considered as aligned along the Oz -axis. Let's consider how the spherical image, $\mathbf{M}_S(\theta, \varphi)$ can be transformed when the robot moves. The rotation of the

robot (at the same position) around its vertical axis by angle φ induces the rotation of the sphere around the Oz axis. Using an omnidirectional camera is an advantage to have a complete view so that the image content information does not change when changing the orientation. We exploit the symmetry of revolution of the considered kernel supports to remove the integration over the rotation group action.

When the robot motion involves a translation (robot changes in position), the transformation acting onto the spherical image is composed of translations in θ and φ . We assume these transformations still verifying a group action, as almost scene information remain present in the omni-image. Due to the non uniform resolution and distorsion of omnidirectional images (see fig. 1 for an example), the image point transformations are not uniform and are weighted by the variation of Haar measure $dg = \sin\theta d\theta d\varphi$. The group parameters acting on the images are translation parameters in $\theta \in [\pi/2, \pi[$ and $\varphi \in [0, 2\pi[$.

3.4 The Kernel Function

The averaging technique to construct invariant features depends on kernel function f . The definition of kernel f appears to be important for the robustness and stability of the built invariant features. For instance, you can define $f(\mathbf{M}_S) = \mathbf{M}_S(0,0)$ as a kernel, that just takes the grey level of the pixel. Evaluating the Haar integral can thus be interpreted as a group averaging of the image brightness information. However, the descriptive power of the mean value is poor when compared to a histogram for example. In [25,10], the authors use a large set of monomials and relational kernel functions with local support to increase the completeness and non ambiguity of the invariant sets.

We use a kernel function of local characteristics based on a Difference of Gaussians (*DoG*) (eq.3). The *DoG* is usually applied for keypoint detection, and is shown [19] to be a good approximation for the σ^2 -Gaussian Laplacian which is invariant² to affine change of luminosity, rotation and locally invariant to perspective transform.

$$DoG(\theta, \varphi, \sigma) = \frac{1}{2\pi(k\sigma)^2} e^{-\frac{\theta^2 + \varphi^2}{2(k\sigma)^2}} - \frac{1}{2\pi\sigma^2} e^{-\frac{\theta^2 + \varphi^2}{2\sigma^2}} \quad (3)$$

Kernel function $f : \mathbf{M}_S \mapsto \mathcal{C}_f$ defines a mapping from spherical image \mathbf{M}_S to feature space \mathcal{C}_f . $f(\mathbf{M}_S)$ denotes the image of local features obtained by Haar integration and is produced by: 1) convoluting the (grey-scaled) spherical image points with the difference of two nearby scale gaussians; 2) averaging on the pixel neighbors belonging to DoG support Δ_{DoG} (of size $6k\sigma$); and 3) partitioning the DoG space into a fixed partition $\{DoG_i\}_{i=1,\dots,k}$. Similarly to [25], we build fuzzy partitions using continuous triangle functions to avoid discontinuities of feature assignments at the edges of DoG supports. We thus produce

² [17] referenced by Lowe [19].

$f(\mathbf{M}_S) = \{f_i(\mathbf{M}_S)\}_{i=1,\dots,k}$ that we normalize to make the sum at a given point of the f_i feature space equals one³.

$$f(\mathbf{M}_S(\theta_0, \varphi_0)) = \left\{ \sum_{(\theta, \varphi) \in \Delta_{D \circ G_i}} D \circ G_i(\theta - \theta_0, \varphi - \varphi_0, \sigma) * \mathbf{M}_S(\theta, \varphi) \right\}_{i \in [1, 2, \dots, k]} \quad (4)$$

Haar integration consists then in a course (path) between the *feature* image points belonging to every f_i , weighted by the Haar measure depending on their position in the image: $I_{Haar} = \{I_{Haar_i}\}_{i=1,\dots,k}$ (eq.5). Written in the discrete case, we have for $i = 1, \dots, k$:

$$I_{Haar_i}(\mathbf{M}_S) = \frac{1}{\Delta_\theta} \frac{1}{2\pi} \sum_{\theta \in \Delta_\theta; \varphi \in [0, 2\pi[} f_i(\theta, \varphi) \sin \theta d\theta d\varphi \quad (5)$$

This distribution looks like a histogram, $h(I_{Haar}(\mathbf{M}_S)) = \{I_{Haar_i}(\mathbf{M}_S); i = 1..k\}$ of invariant features and constitutes the image signature (parameterized by σ).

3.5 The Similarity Measure of Distributions

Images taken at close locations are likely to be very similar. Finding the most similar reference image to the current image taken by the robot needs a distance measure to compare distributions derived from the images. This issue has been thoroughly discussed in the literature, from simple bin-by-bin measures (e.g. quadratic form distance) to more complex measures like Minkowski form distances, histogram intersection, chi-square statistic, Jeffrey divergence or earth movers distance (EMD).

In a previous work [6], we have tested different similarity measures and the EMD distance performs better for our image database. In this paper, and in order to compare our method to other signature extraction techniques [23,21], we use the same similarity measure as in the cited works, i.e. L_1 -norm similarity measure.

4 Comparative Study and Experimental Setup

In this section, we compare our method of Haar invariant signatures to related methods of signature extraction: The zero phase representation developed by Pajdla (ZPR,[23]) and representation using Fourier coefficients of an unwarped image suggested by Menegatti [21].

4.1 A Brief Description of the Compared Methods

Menegatti’s method starts by unwrapping the raw omnidirectional image into a cylindrical image. It relies on a Fourier decomposition along the lines of the

³ f_i could be seen as the probability for a characteristic to belong to a given feature bin.

cylindrical image so as to construct a representation that is invariant to rotations. The descriptive power of the representation is set by adjusting the number of Fourier coefficients used in the representation. The higher the number of coefficients, the more accurate the representation. The last coefficients correspond to high frequencies whereas first coefficients correspond to low frequencies. This method tends to allow a greater importance to low frequencies as the number of coefficient diminishes, the representation thus being less accurate.

The Zero-Phase representation also allows to cope with rotation effects on the image representation. It relies on a correlation measure between images which is invariant to rotations of the sensor (along the optical axis). As in Menegatti’s solution, the raw image is unwrapped onto a cylinder. Then, the 2-D Fourier transform is calculated. The phase of the Fourier transform is shifted so as to get the phase of the first component equal to zero. An inverse Fourier transform is then applied to get back into the original space.

4.2 The Experimental Setup

In this comparative study, we implement the preceding signature extraction methods and test their robustness in indoor environments, composed of five rooms in our lab, where the perceptual aliasing is particularly high.

The Koala robot, endowed with a catadioptric camera (RemoteReality), was moved in the environment and we have built an image database composed of approximately 250 images evenly distributed in the five rooms. Examples of the images in the base are shown in Fig. 1. Images taken in a given room were manually clustered in an associated directory, thus allowing to calculate the *centroid* and the *variance* of the cluster with respect to the L_1 -norm. This was repeated for every room the robot has surveyed.

To assess the classification performance of the compared methods, we use the *Fisher criterion* which measures the separation between two classes. It is defined as $\mathcal{J} = (\eta_1 - \eta_2)^2 / (\sigma_1^2 + \sigma_2^2)$, evaluating the ratio of the squared distance between centroids η_1 and η_2 of the classes over variances σ_1^2 and σ_2^2 of the representations belonging to them.

The higher this ratio is, the better the separability between two classes is. We assess the categories produced by the different methods in the light of this criterion. The best representation will be that with the highest ratio. We also want to show that our method benefits from interesting properties. Namely, its accuracy can easily be tuned using one parameter, the scale-variable $k\sigma$ (fig. 3(b)).

5 Comparative Experimental Results

The experimental results for the robot position estimation, using Zero-Phase representation and Fourier low-frequency coefficients techniques were compared with the results obtained applying Haar integral signature extraction, using their Fisher criteria.

5.1 The Discrimination Performance

To differentiate between different places in the lab, in spite of the high perceptual aliasing, we need good clustering properties of the extracted signatures. Figure 3 shows the comparative results when the precedent signature extraction techniques were used.

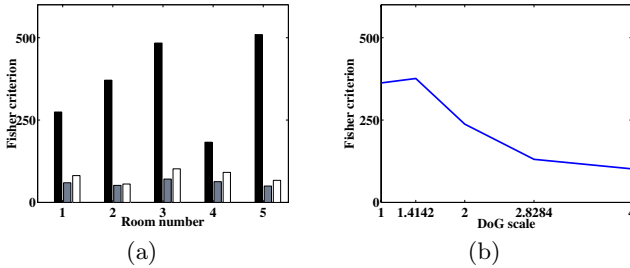


Fig. 3. Left: Comparison of the Fisher criterion of the three tested methods for the five rooms explored by the robot. *Black*: Haar invariant signatures ($\sigma = 1$ and $k = \sqrt{2}$), *grey*: ZPR and *white*: Fourier transform. Right: Evolution of the Fisher criterion for different scales of the *DoG* used.

As expected, we obtain significantly better results for places discrimination than with ZPR or Fourier transform representations, because our method takes account of the geometrical and projective properties of the sensor. Unlike the ordinary histograms, Haar integral distributions have the advantage of capturing the local structure held in the image when the images are transformed, by weighting the distributions by the Haar measure dg . This advantage is preserved at different scales of the kernel function as shown in fig. 3(b).

The different techniques implemented have equivalent computational complexity: $\mathcal{O}(n \log(n))$ for the Fourier transforms and $\mathcal{O}(n \cdot m)$ for the kernel convolution introduced by our method, where n is the image size and m the kernel one. On an AMD 1800MHz, the whole process of the signature construction takes approximatively 0.30s. The comparison process between a signature and all the signatures (250) of the image database takes around 400 μ s.

5.2 The Position Estimation Performance

We are now interested to evaluate the signature discrimination for the robot position estimation inside each room. When the robot moves in an area, small distortions are produced in the omnidirectional image space and we should expect small changes in the feature space. This property is characterized [4] by the *continuity* of the used transform for signature extraction, with respect to a certain metric.

In this experiment, the robot was moved in every room of the lab and we have extracted signatures at different locations using the previous methods. As is

suggested by the Fisher criterion, the variation among the images of a given room is small for our method when compared to the distance between the centroids of two rooms. Our method produces signatures that are partial integral invariants which means they are good local invariants, i.e. for reasonable movements they should not vary a lot. We study this through the evolution of the L_1 -norm between signatures of the image database with respect to the physical distance (in meters) between the positions they are associated to. As the signatures of the different methods presented here belong to different feature spaces, the produced L_1 -norm are not directly comparable (up to five orders of magnitude). Thus, the L_1 -norms have been normalized by an affine change of variable to bring the L_1 -norm in a given room, in $[0, 1]$. The results are shown in fig. 4. The Haar invariant signatures produces smaller variations than the ZPR or Fourier coefficients. This is not surprising as we were looking to produce an invariant. But an interesting fact is that the variation of the Haar signatures is still *monotonic*. Thus, we can define a bijective relation between the L_1 -norm and the physical distance, allowing a localization inside a given room using these signatures.

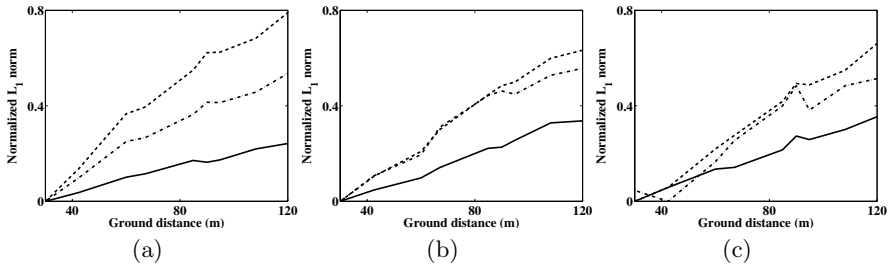


Fig. 4. The evolution of the L_1 -norm according to the distance on the ground (in meters) for different rooms ((a) \rightarrow (c)) applying the studied signature extraction techniques. *Solid line*: Haar invariant distributions, *dot-dashed line*: Fourier transform, *dashed line*: Zero-phase representation.

6 Conclusion and Ongoing Work

We have proposed an efficient methodology to build invariant signatures for omnidirectional image based localization applications. The Haar integral formalism offers a solid theoretic foundation to the invariant signatures of images we have introduced. Our method benefits from the local invariance properties of the defined kernel function and the global invariance of Haar integration. During the integration process, we introduce the geometric and projective transformations of omnidirectional sensors, as produced by the robot movements.

We have compared our method to different techniques of signature extraction, for omnidirectional image based localization: Zero-phase representation, Fourier transform signatures and histograms of local characteristics. Our method proved to get significantly better results for indoor environment recognition and robot

localization. Using the Fisher criterion, the built signatures have figured out a wide separation ability of room classes, contributing to reduce the perceptual aliasing. Moreover, the smooth variation and the continuity property of the built Haar signatures, inside each category, provides a good approximation to the robot position for localization.

Additional development is under way to optimize the exploration training stage and automatic recognition of the robot position by using additional knowledge on the local robot movements in the Haar integral. We are also exploring different methods to build kernel functions as we believe that this will influence the stability and the precision of localization in a positive way.

References

1. H. Aihara, N. Iwasa, N. Yokoya, and H. Takemura. Memory-based self-localisation using omnidirectional images. In Anil K. Jain, Svetha Venkatesh, and editors Brian C. Lovell, editors, *14th International Conference on Pattern Recognition*, volume I, pages 1799–1803, 1998.
2. H. Andreasson, A. Treptow, and T. Duckett. Localization for mobile robots using panoramic vision, local features and particle filter. In *IEEE International Conference on Robotics and Automation*, 2005.
3. J. Borenstein, B. Everett, and L. Feng. *Navigating mobile robots: Systems and techniques*. A. K. Peters, Ltd., Wellesley, MA, February 1996.
4. H. Burkhardt and S. Siggelkow. *Nonlinear Model-Based Image/Video Processing and Analysis*, chapter Invariant features in pattern recognition - fundamentals and applications, pages 269–307. John Wiley & Sons, 2001.
5. J. A. Castellanos and J. D. Tardos. *Mobile robot localization and map building: A multisensor fusion approach*. Kluwer Academic Publishers, Boston, Mass, 2000.
6. C. Charron, O. Labbani-Igibida, and E.M. Mouaddib. Qualitative localization using omnidirectional images and invariant features. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, Edmonton, Canada, August 2005. IEEE/RSJ.
7. J. Gaspar, N. Winters, and J. Santos-Victor. Vision-based navigation and environmental representations with an omnidirectional camera. *IEEE Transactions on Robotics and Automation*, 16(6):890–898, December 2000.
8. C. Geyer and K. Daniilidis. Catadioptric projective geometry. *International Journal of Computer Vision*, 45(3):223–243, 2001.
9. J. Gonzalez and S. Lacroix. Rover localization in natural environments by indexing panoramic images. In *IEEE International Conference on Robotics and Automation*, pages 1365–1370, 2002.
10. A. Halawani and H. Burkhardt. Image retrieval by local evaluation of nonlinear kernel functions around salient points. In *International Conference on Pattern Recognition*, volume 2, pages 955–960, August 2004.
11. A. Halawani and H. Burkhardt. On using histograms of local invariant features for image retrieval. In *IAPR Workshop on Machine Vision Applications*, pages 538–541, May 2005.
12. H. Ishiguro and S. Tsuji. Image-based memory of environment. In *IEEE/RSJ International Conference on Intelligent Robots and Systems*, volume 2, pages 634–639, 1996.

13. M. Ishikawa, S. Kawashima, and N. Homma. Memory-based location estimation and navigation using bayesian estimation. In *International Conference On Neural Information Processing*, volume 1, pages 112–117, October 1998.
14. H. Burkhardt J. Wolf, W. Burgard. Using an image retrieval system for vision-based mobile robot localization. In M. S. Lew, N. Sebe, and J. P. Eakins, editors, *Proc. of the International Conference on Image and Video Retrieval (CIVR)*, pages 108–119. Springer-Verlag Berlin Heidelberg, 2002.
15. M. Jogan and A. Leonardis. Robust localization using an omnidirectional appearance-based subspace model of environment. *Robotics and Autonomous Systems*, 45(1), 2003.
16. A. Leonardis and H. Bischof. Robust recognition using eigenimages. *Computer Vision and Image Understanding Special Issue on Robust Statistical Techniques in Image Understanding*, 78(1):99–118, 2000.
17. T. Lindeberg. Scale-space theory: A basic tool for analysing structures at different scales. *Journal of Applied Statistics*, 21(2):224–270, 1994.
18. D. Lowe. Object recognition from local scale-invariant features. In *International Conference on Computer Vision*, pages 1150–1157, September 1999.
19. D. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.
20. S. Maeda, Y. Kuno, and Y. Shirai. Active navigation vision based on eigenspace analysis. In *International Conference on Intelligent Robots and Systems*, pages 1018–1023. IEEE/RSJ, 1997.
21. E. Menegatti, T. Maeda, and H. Ishiguro. Image-based memory for robot navigation using properties of the omnidirectional images. *Robotics and Autonomous Systems, Elsevier*, 47(Issue 4):251–267, 2004.
22. E. Menegatti, M. Zoccarato, E. Pagello, and H. Ishiguro. Image-based monte-carlo localisation with omnidirectional images. *Robotics and Autonomous Systems, Elsevier*, 48(Issue 1):17–30, 2004.
23. T. Pajdla and V. Hlaváč. Zero phase representation of panoramic images for image based localization. In F. Solina and A. Leonardis, editors, *8-th International Conference on Computer Analysis of Images and Patterns*, pages 550–557, Ljubljana, Slovenia, September 1999. LNCS, Springer Verlag.
24. H. Schulz-Mirbach, H. Burkhardt, and S. Sigglekow. Using invariant features for content based data retrieval. In *Workshop on Nonlinear Methods in Model-Based Image Interpretation*, pages 1–5, Lausanne, Switzerland, September 1996.
25. S. Sigglekow. *Feature histograms for content-based image retrieval*. PhD thesis, Universitat Freiburg im Breusgau, 2002.
26. H. Tamimi, H. Andreasson, A. Treptow, T. Duckett, and A. Zell. Localization of mobile robots with omnidirectional vision using particle filter and iterative sift. In *European Conference on Mobile Robots*, Ancona, Italy, 2005.
27. I. Ulrich and I. Nourbakhsh. Appearance-based place recognition for topological localization. In *International Conference on Robotics and Automation*, pages 1023–1029, San Francisco, CA, April 2000. IEEE.

Accurate 3D Structure Measurements from Two Uncalibrated Views

Benjamin Albouy¹, Emilie Koenig², Sylvie Treuillet², and Yves Lucas³

Vision and Robotics Laboratory, Orleans University, France

¹ ENSI, 10 Bd Lahitolle 18000 Bourges cedex 2

Benjamin.Albouy@ensi-bourges.fr

² Polytech'Orléans, site Galilée, 12 rue de Blois, BP 6744

45067 Orléans

{Sylvie.Treuillet, Emilie.Koenig}@univ-orleans.fr

³ IUT Mesures Physiques,

63 Avenue De Lattre de Tassigny 18020 Bourges cedex

Yves.Lucas@bourges.univ-orleans.fr

Abstract. We have developed an efficient algorithm to compute an Euclidean reconstruction from only two wide-baseline color images captured with a hand-held digital camera. The classical reconstruction scheme has been improved to boost the number of matches by a hierarchical epipolar constraint during an iterative process and an ultimate step of dense matching based on affine transformation. At the output, between three to four thousands points are reconstructed in 2 minutes on 1024x768 images. The stability of the algorithm has been evaluated by some repetitive tests and the quality of the reconstruction is assessed according to a metric ground truth provided by an industrial 3D scanner. The averaged error on 3D points is around 3.5% reported to the model depth. Such a precision makes this technique suitable for wound volumetric assessment in clinical environments using a hand held digital camera.

1 Introduction

Dense two frame stereo matching techniques for 3D reconstruction have been widely studied under known camera geometry [1]. A common drawback is the requirement of a calibration step each time the field of view has to be changed, as it affects the configuration of the two cameras. Another problem is that small baseline degrades the triangulation step and results in low precision data. Self-calibrated vision techniques have then been introduced to compute structure from motion with images as the only input [2]. Self-calibration allows to update the camera parameters modified by zooming or focus operations without the use of costly calibration pattern. Most of the works consider large sequences of images combined by an optimization process for object 3D capture, augmented reality [3,4] and robotics applications [5]. Because of the overlap between consecutive frames, the matching is then easier to realize between wide baseline views which is a key point for self-calibration stability and triangulation accuracy.

Conversely, our approach is based on a very small set of images captured from very different points of view with a hand-held digital camera [6]. The great advantage of this approach lies in the flexibility of the image capture to produce a 3D textured model. But behind the easy and free image capture, real technical difficulties are dissimulated due to wide-baseline, such as strong scale changes in texture or varying lighting conditions. Wide-baseline matching has been a tricky problem for a long time, but recent works demonstrate the robustness of some local descriptors in this case. A comparative study [7] on matching and object recognition clearly confirms the robustness and the distinctive character of the SIFT descriptor. This scale and affine-invariant region descriptor is presented in [8]. PDE-based approach for dense reconstruction from multiple wide-baseline views has also been developed [9] with an acceptable computational time (around 15 minutes).

This paper presents important enhancements to our earlier work [6]. As a result, we are now able to reconstruct between three to four thousands 3D points from only a pair of color images (1024 x 768 pixels) in less than 2 minutes. This multiplication of the robust matches enables a precise 3D reconstruction, validated against a ground truth provided by a commercial 3D scanner. Two major improvements have been introduced in the reconstruction chain: a decreasing bandwidth of the epipolar constraint to combine a higher outlier rejection with the fundamental matrix convergence during an iterative process and a local mapping of triangles for real dense matching. Increasing the number of robust matches results in higher accuracy and 3D reconstruction stability [12].

The accuracy assessment of the Euclidean reconstruction is essential for volumetric measurement applications. Inside a reconstruction pipeline, two errors are traditionally observed by the authors: the residual error measuring the average distance between matched points and the associated epipolar lines, and the reprojection error. However, these errors are estimated in the image plane and do not reflect the final 3D structure accuracy. The precision of the inferred 3D model is generally not assessed in a quantitative manner but only by a visual inspection [2,9]. We propose to evaluate the metric reconstruction error according to a ground truth given by an industrial 3D scanner.

On the other hand, very few works consider the instability of the stochastic methods generally used for the fundamental matrix estimation [10,11,12]. Running the algorithm a second time on the same image pair will not give the same 3D structure because of the stochastic outliers rejection. The error led by the fundamental matrix estimation is particularly large near some singular cases [11,12]. We propose to address this issue by some repetitive tests to evaluate an error confidence interval.

The paper is organized as follows. Section 2 describes the software architecture of the reconstruction chain. In section 3 we develop the two steps introduced to boost the number of matches. Experimental results are presented in section 4 and section 5 concludes the paper.

2 Reconstruction Chain

The reconstruction pipeline generally adopted is composed of 3 stages: matching, self calibration, triangulation and bundle adjustment (Fig.1). Our wide-baseline

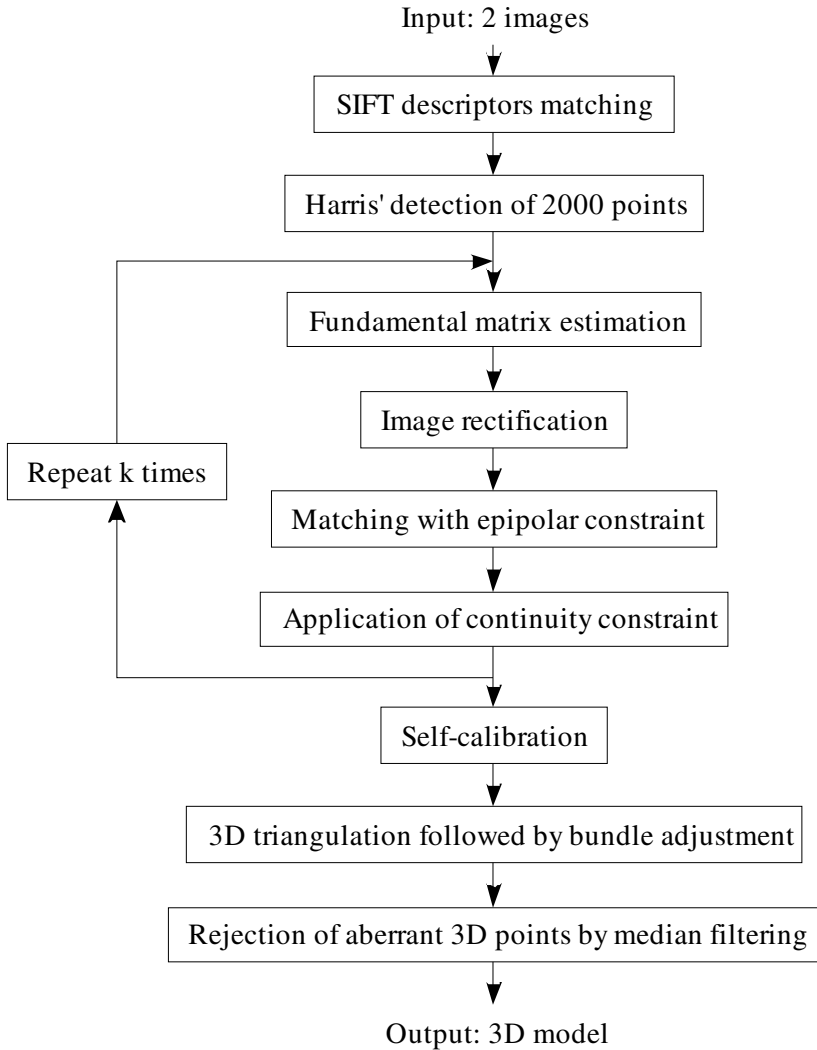


Fig. 1. Reconstruction pipeline

stereo reconstruction starts with the matching of some invariant singular points using SIFT descriptor [8] on a pair of wide-baseline color images.

Fig.2 shows an example of twenty initial matches automatically extracted. The colored pins placed in the field of view provide a metric reference and are also included as two reliable matches. These robust matches are used to compute the initial estimation of the fundamental matrix thanks to LMedS method [10]. About two thousands color interest points are detected by the Harris detector adapted by Gouet [13]. These points are then matched by applying an iterative process. During iterations, the images are rectified [14] to use epipolar constraint for matching. Once the fundamental

matrix is estimated and the images are rectified, the search for corresponding points between the views may be reduced to a small band centered around the epipolar line. Matching is based on the cross “Winner Take All” algorithm with a ZSAD similarity score under a continuity constraint. This constraint verifies that the displacement vector does not differ too much from the average vector calculated on a circular neighbourhood of 50 pixels. The difference between the two vectors must be less than 25% in magnitude and less than $\pi/10$ in argument. Once a set of point matches has been created, an improved estimate of fundamental matrix can be calculated between the pair of views, and so on. The improved fundamental matrix makes the correspondence more efficient and improves the outliers rejection by the well-established and robust LMedS method [2]. The final fundamental matrix is computed using all the inliers and matches are triangulated in 3D-space. A bundle adjustment algorithm [15] is applied to optimize the location of all 3D points and the camera parameter (focal length). Next, a smoothed triangular mesh is constructed from the resulted points cloud by applying a median filter. The first image is then mapped as a texture on the mesh.

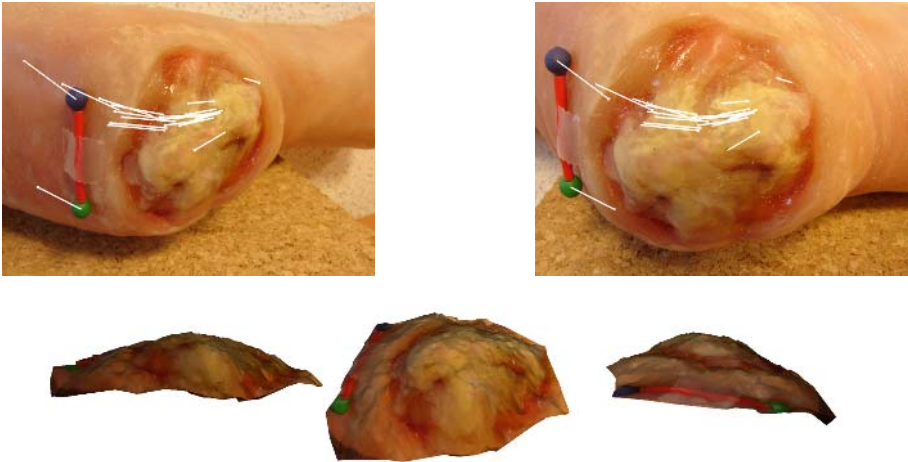


Fig. 2. Initial matches inferred by SIFT descriptor (*top*). Three textured views of the reconstruction from two uncalibrated wound images (*bottom*).

At this stage, typically less than 300 points are reconstructed after 7 iterations. The number of inlier matches is not really sufficient for a high-quality mesh and a good metric reconstruction. Furthermore, the relative error on focal length estimation decreases with the number of matches [12]. So to refine mesh, it is necessary to increase drastically the number of right matches, and not ordinary interpolate data points.

3 Multiplying the Number of Matches

To obtain a denser 3D reconstruction, we boost the number of matches by two extensions to the above algorithm, these key improvements are developed in next sections.

The first one introduces a hierarchical epipolar constraint during the iterative process. The second one presents an additional step of matching refinement.

3.1 Hierarchical Epipolar Constraint

The main idea is that an iterative estimation of the epipolar geometry should be more efficient if stronger constraints are applied at each step: more outliers will be rejected and a small number of matching candidates makes the cross picking of good matches easier (maxima of correlation score). A first improvement is obtained by easily replacing the constant band width usually used in the epipolar constraint by a decreasing width during iterations. The bandwidth has an important effect on the number of matches selected by the “Winner Takes All” algorithm. So, by progressively limiting the bandwidth, the number of matches increases without degrading the quality. The efficiency of this decreasing step is illustrated on fig. 3.

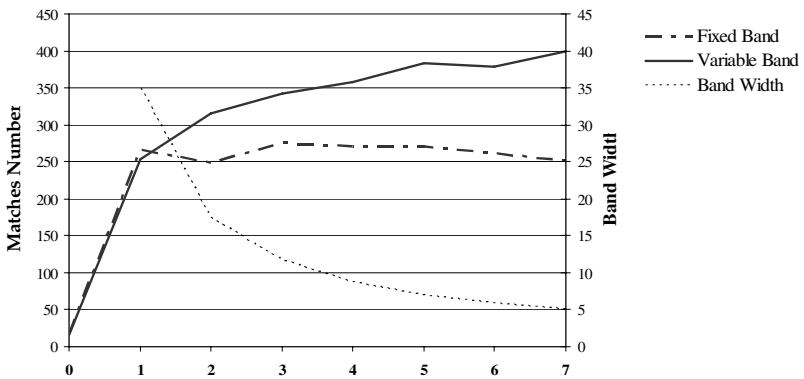


Fig. 3. Number of matches during iterations with a constant bandwidth and with a decreasing bandwidth for the epipolar constraint

The band width (dot line) follows a linear function inversely proportional to the loop index, decreasing from 35 to 5 pixels on both sides of the epipolar line. Using a constant width (dashed-dot line) leads to a stable number of matches after one iteration; on the contrary, a variable bandwidth (bold line) boosts the number of matches around 60% after seven iterations.

3.2 Affine Transformation of Similar Triangles

An iterative process to extract consistent matches with fundamental matrix estimation provides a limited set of inliers (typically around four hundreds pairs). So, an ultimate step is introduced at the output of our algorithm to multiply the number of matches in the pair of original images. A classical interpolation would only create additional data points without changing the precision of the 3D reconstruction. On the

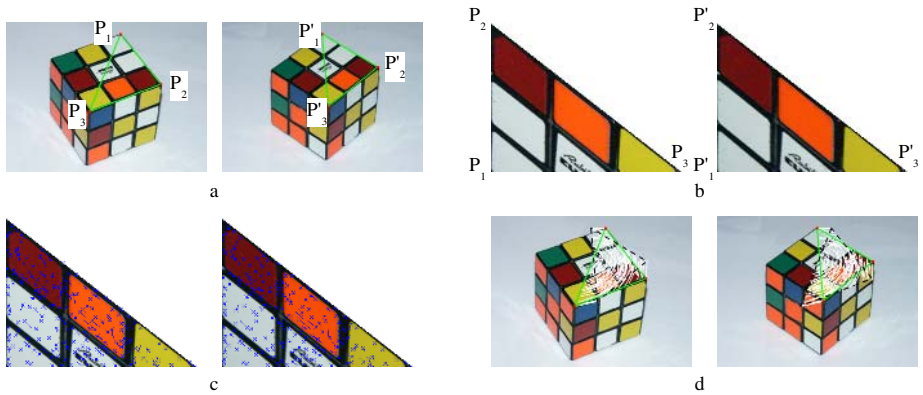


Fig. 4. An example of matches multiplication by applying an affine transformation on homologous triangles. (a) Triangular region computed in the first image and reported in the second one. (b) Affine transformation applied to triangles. (c) Harris corners extraction. (d) Matching reported on original triangles.

other hand, using a similarity score is not efficient since the images are taken from very different viewpoints. So, the idea is to apply a triangulation between all the inliers extracted in the two images and then to bring them locally closer by applying an affine transform. To do this, the Delaunay triangulation computed on the matched points in the first image (P_1, P_2, P_3), is reported in the second one (P'_1, P'_2, P'_3). Then the template affine transformation is defined by the three matched vertexes and applied to triangles to put them in the same frame. The corners extracted by the Harris detector inside each transformed triangular region are then matched with the ZSAD similarity score controlled by a relaxation algorithm. We use the relaxation method proposed by Zhang and reported in [13], which favour matches that have numerous symmetric matches in a neighbourhood representing $1/8$ of the image size. An example of our method is given in fig.4. This dense matching provides 3000 to 4000 matched pairs and allows a more precise reconstruction.

4 Results

We have tested the improved algorithm on several pairs of real images. Results are divided in two sections. The first deals with some repetitive tests realized on the same image pair to illustrate the stability of the algorithm. In the second one, we compare the 3D structures inferred with a very precise 3D model of the object given by an industrial 3D scanner.

4.1 Repetitive Test on Real Images

To test the stability of the reconstruction algorithm, we run it 10 times on the same image pair presented in fig.2. These color images show an artificial heel wound designed by a professional of special effects. These color images present a wide-

baseline with a strong scale change. They have been captured by a hand-held digital camera (Sony DSC-H1) in a macro mode (fixed focal length) with a large image size (2048 x 1536 pixels), and a high quality JPEG compression. The images are then resized to 1024x768 for current processing. To avoid the singular orbital configuration at the self-calibration stage [12], the two points of view are separated by an angle around 30 degrees and a 1.5 ratio on the distances relative to the object. Ten reconstructions are computed from the same initial matches (fig.2). Because of the stochastic nature of the algorithm, each 3D structure inferred is different to the others. After the fundamental matrix estimation and outliers rejection, the average of the residual errors is equal to 1.28 pixels, with a standard deviation of 1.05 pixels on the ten tries. After the bundle adjustment, the average of the reprojection errors is 0.92 pixel, with a standard deviation 0.77 pixel. This error is not high compared to some published shape-from-video results obtained from large video sequences (typically 50 to 150 frames). Computation time depends on the number of reconstructed points with the image size, but is generally between 2 and 3 minutes for 3000 to 4000 reconstructed points on a Pentium IV 3.4 GHz based computer. A software optimization could reduce it by a half.

4.2 Accuracy Assessment of 3D Structure

The accuracy of the final 3D structures has been evaluated by comparing them with a ground truth provided by an industrial 3D scanner. An ordinary visual inspection with a flattering texture mapping does not yield a quantitative evaluation. The ground truth data has been provided by KREON Technologies ZEPHYR[®] system, based on active triangulation: the object is scanned by a hand-held scanner that projects a red laser line seen from a fixed angle by the camera. The measuring head displacements are recorded by an articulated robot arm. About 500 points are extracted on the laser line. The system resolution is 3 μm with reproducibility of 9/15. For our object, the reference surface generated by this scanner contains 41651 points in the measured area.

To compare our reconstruction with this ground truth model, a scaling and a rigid registration of the 3D data is necessary. This has been done by using 3DReshaper[®] software distributed by TECHNODIGIT. The global registration is composed of two parts. The first step is a manual initialization of the 3D-3D rigid registration. To do this, six remarkable points are chosen manually on the cloud and matched with their equivalent on the reference surface. The distances between the points are minimized to initialize the registration. Then, an optimization is processed using the Iterative Closest Point algorithm [16]. Finally, the coordinates of 3D registered points and those of their orthogonal projection on the reference surface are recorded in an Excel file.

The comparison between the registered cloud and the ground truth model is then evaluated by the residual deviations on each 3D point: $DevX_i$, $DevY_i$ and $DevZ_i$ represent the signed distances on each direction between a point and the reference surface along the normal estimated on adjacent triangles (orthogonal projection). We observe three types of error (Table 1): the signed residual error, the averaged distance and the root mean squared error. The first one is directly provided by the 3DReshaper software to evaluate the registration convergence and error distribution. The distributions of the signed local deviations are presented on Fig.5 by their confidence intervals at 95%

(average \pm double standard deviation) for the ten repetitive reconstructions obtained on the same image pair (fig.2). Despite a relative instability, they are almost zero centered for all reconstructions. Fig.6 illustrates two examples of error maps observed on the 3D surface with the corresponding distribution color scale to the left.

The averaged distance on all 3D points is based on the Euclidian distance:

$$AVD = \frac{1}{N} \sum_{i=1toN} \sqrt{DevX_i^2 + DevY_i^2 + DevZ_i^2} \tag{1}$$

where N is the number of 3D points considered.

The classical root mean squared error formula is:

$$RMS = \sqrt{\frac{1}{N} \sum_{i=1toN} (DevX_i^2 + DevY_i^2 + DevZ_i^2)} \tag{2}$$

Table 1 gives the global accuracy evaluation according to *AVD* and *RMS* defined above over the ten repetitive tests. The *RMS* is around one millimeter with a maximum error of 1.83. Considering the averaged distance observed on 3D points, it is less than one millimeter for all reconstructions and below half a millimeter for six of them. The relative error in percent is reported to the model depth ($Y_{max}=17.23$ mm). It also interesting to note that the average relative value 3.5 % is close to the one evaluated on simulated data presented in preliminary results [6,10]. These results are good if we consider the difference of clouds density: our reconstruction presents a number of 3D points ten times smaller than the reference surface. The manual initialization of the registration raises a repeatability problem. The registration has been realized 20 times on the same pair of 3D data. From these repetitive tests, the average of the ten values for *AVD* is 0.47 mm with a standard deviation of 0.03 mm, e.g. 6% of the average. We conclude that the manual pairing used to initialize registration doesn't really influence the final result.

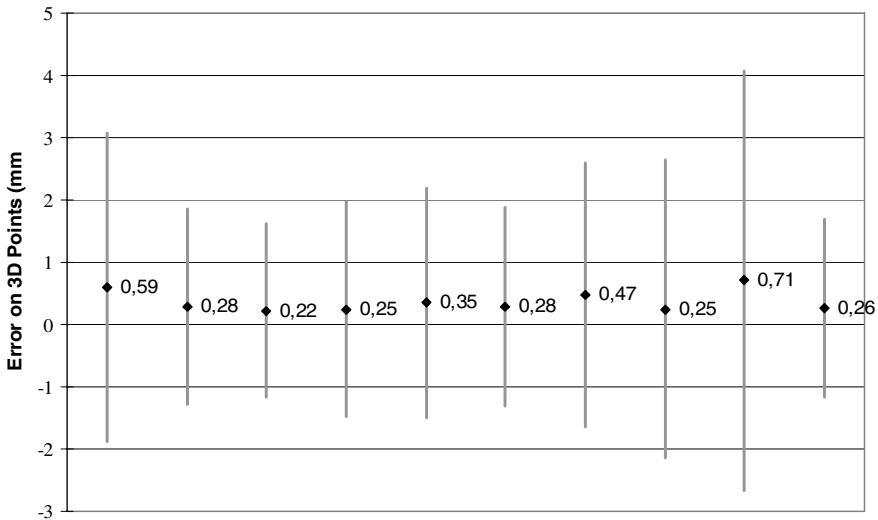


Fig. 5. Confidence intervals at 95% for the local signed errors between 3D points and ground truth surface for the ten repetitive reconstructions from the same image pair

Table 1. Statistics of the observed errors over the ten repetitive tests on the same image pair. Relative error is the AVD reported to the model dimension.

	Signed Error (mm)	RMS (mm)	AVD (mm)	Relative Error (%)
Average	0,37	1,06	0,63	3,65
Standard deviation	0,17	0,34	0,18	1,05
Min value	0,22	0,73	0,44	2,56
Max value	0,71	1,83	0,94	5,44

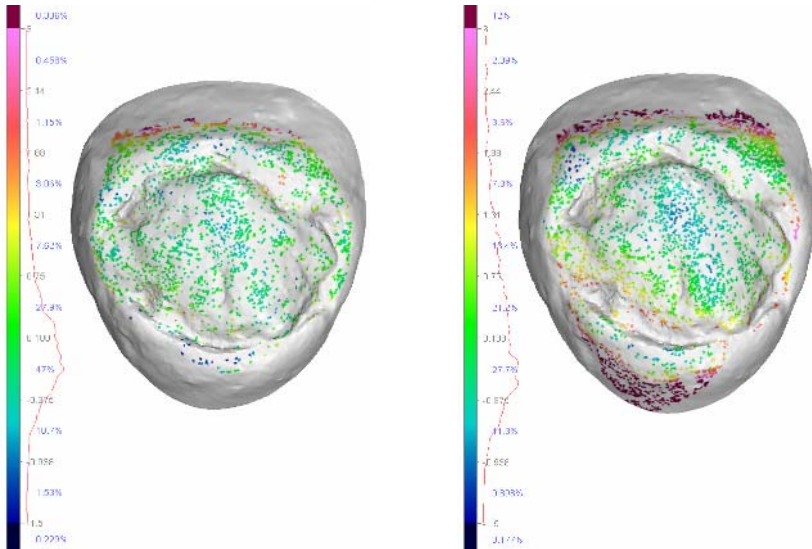


Fig. 6. Error distribution on the 3D surface for the best (*right*) and worst (*left*) reconstruction

5 Conclusion

We have developed an efficient algorithm to compute an Euclidean reconstruction with two uncalibrated wide-baseline images including important scale changes as the only input. The main contribution of the presented work is the mix of existing algorithms into an integrated framework, through which a fully automated reconstruction can be evaluated with respect to a metric ground truth. Two key improvements have been introduced to the common reconstruction pipeline to boost the number of matches. Firstly a hierarchical epipolar constraint is introduced during the iterative process to reject the outliers and to refine the fundamental matrix. And an ultimate step of quasi dense matching based on an affine transformation of homologue triangles in the two images is added at the output of the pipeline to refine the resulting mesh. These improvements provide a dense reconstruction: 3000 to 4000 points from only a pair of images. The stability of the algorithm has been evaluated by some repetitive tests and the quality of the reconstruction is assessed according to a metric

ground truth given by an industrial 3D scanner. The averaged error on 3D points is generally less than millimeter, e.g. a relative error around 3.5% reported to the model depth. In the context of wound volumetric measurement [6], this precision is sufficient for therapeutic following. Based on a free hand-held digital camera, this user-friendly and low cost technique can be widely spread to the medical staff to provide high-quality 3D model.

Acknowledgements

The ESCALE research program is supported by the European Social Funds, the French Delegation of Research and Technology and the region Centre. Thanks to TECHNODIGIT company for free access to 3D Reshaper software and to KREON Technologies for graceful 3D model scanning.

References

1. Scharstein R, Szeliski R.: A taxonomy and evaluation of dense two-frame stereo correspondance algorithms, *Int. J. of Computer Vision*, 47-1/2/3 (2002) 7-42
2. Hartley R.I and Zisserman A.: *Multiple View geometry in Computer Vision*, Cambridge University Press (2000)
3. Pollefeys M., Koch R., Van Gool L.: Self-calibration and metric 3D reconstruction in spite of varying and unknown internal camera parameter, in *Proc. of Int. Conf. on Computer Vision* (1998) 90-95
4. Pollefeys M., Van Gool L., Vergauwen M., Verbiest F. Cornelis K. and Tps J. : Visual modeling with a hand-held camera, in *Int. J. Computer Vision*, 53-3 (2004) 207-232
5. Royer E., Lhuillier, M. Dhome, M. Chateau T.: Localization in urban environment: monocular vision compared to GPS sensor, in *Proc. of Conf. on Computer Vision and Pattern Recognition* (2005) 114-121
6. Albouy B., Treuillet S., Lucas Y., Pichaud J.C.: Volume estimation from two uncalibrated views applied to wound measurement, *Int. Conf. of Image Analysis and Processing, Cagliari* (2005) 945-952
7. Mikolajczyk K. and Schmid C.: Performance evaluation of local descriptors, *IEEE Trans. on PAMI*, Vol.27-10 (2005) 1615-1630
8. Lowe D.: Distinctive image feature from scale-invariant keypoints, *Int. Journal of Computer Vision*, Vol 2-60 (2004) 91-110
9. Strecha C., Tuytelaars T., Van Gool L.: Dense matching of multiple wide-baseline views, *IEEE Int. Conf. on Computer Vision* (2003) 1194-1201
10. Albouy B., Treuillet S., Lucas Y., Birov D.: Fundamental matrix estimation revisited through a global 3D reconstruction framework, *ACIVS, Brussels* (2004) 185-192
11. Albouy B., Treuillet S., Lucas Y.: Mesure volumétrique d'escarres à partir de vues stéréoscopiques non calibrées, *GRETSI, Louvain-la-Neuve* (2005)
12. Sturm P., Cheng Z.L., Chen P.C.Y., Poo A.N: Focal length calibration from two views: method and analysis of singular cases, *Computer Vision and Image Understanding*, Vol. 99-1 (2005) 58-95
13. Gouet V.: Mise en correspondance d'images en couleurs, application à la synthèse de vues intermédiaires, Thèse de Doctorat de l'Université de Montpellier II (2000)

14. Oram D.: Projective Reconstruction and Metric Models from Uncalibrated Video Sequences, PhDThesis, University of Manchester (2001)
15. Lourakis M.I.A., and Argyros A.A.: The Design and Implementation of a Generic Sparse Bundle Adjustment Software Package Based on the Levenberg-Marquardt Algorithm, Institute of Computer Science – FORTH, Heraklion, Greece, n°340 (2004)
16. Besl P.J., and Mc Kay N.D.: A method for registration of 3D shape, IEEE Trans. on PAMI, Vol.14-2 (1992) 239-256

A Fast Offline Building Recognition Application on a Mobile Telephone

N.J.C. Groeneweg, B. de Groot, A.H.R. Halma, B.R. Quiroga,
M. Tromp, and F.C.A. Groen

University of Amsterdam
Informatics Institute
Kruislaan 403
1098 SJ Amsterdam

{njgroene, bgroot, ahalma, bquiroga, mtromp, groen}@science.uva.nl

Abstract. Today most mobile telephones come equipped with a camera. This gives rise to interesting new possibilities for applications of computer vision, such as building recognition software running locally on the mobile phone. Algorithms for building recognition need to be robust under noise, occlusion, varying lighting conditions and different points of view. We present such an algorithm using local invariant regions which allows for mobile building recognition despite the limited processing power and storage capacity of mobile phones. This algorithm was shown to obtain state of the art performance on the Zürich Building Database (91% accuracy). An implementation on a mobile phone (Sony Ericsson K700i) is presented that obtains good performance (80% accuracy) on a dataset using real-world query images taken under varying, suboptimal conditions. Our algorithm runs in the order of several seconds while requiring only around 10 KB of memory to represent a single building within the local database.

1 Introduction

In today's society mobile technology plays an important role. With the widespread use of next generation mobile phones a large percentage of the population carries a potent processing unit, which nowadays is accompanied by a digital camera. This setting has presented interesting opportunities for the development of new applications of artificial intelligence, in particular of computer vision. One interesting idea in this direction is an application for a mobile device for the recognition of buildings in an urban environment using only a still camera image as input. Such an application could be used within the tourist industry to provide people with interesting information on photographed buildings, such as opening hours and historical background information, without the need for any additional hardware other than a camera equipped mobile phone.

Although such applications have recently been described within the literature [2] these systems all rely on a client-server architecture in which a user takes a photograph with his mobile device, sends the image to the server which carries

the computational workload and subsequently returns information to the user. Such an architecture has some serious drawbacks. The communication overhead is costly in time and is expensive for the user since mobile network providers usually charge data transfer across their network. Secondly, such an application requires coverage by a mobile network, limiting the application to use within urban environments where coverage is high. A more interesting version of such an application would run locally on the mobile phone without requiring any communication with an external server, assuming the application is acquired on beforehand. However, developing a computer vision application that works in acceptable time on a limited device such as a mobile phone requires a careful selection of techniques, as the mobile platform imposes constraints on both computational power and storage capacity. The application suggested here tackles these practical issues and performs fast offline building recognition in a robust manner.

The aforementioned constraints strongly influence the choice of algorithms and restricts one to a selection of relatively cheap techniques which require as little storage as possible. This implies that great care should be taken to ensure that the representation of buildings in the local database (on the mobile phone) is as compact as possible without affecting the performance of the classification algorithm used. Ideally, one would train a learning algorithm offline and provide the mobile phone with a compact representation of the original database which contains only those image features which are required to correctly classify the buildings in the database. The mobile device would subsequently perform feature extraction/construction on the captured image and feed the result to a compactly represented decision rule, thusly eliminating the need of communication with an external server. On the other hand the application requires an approach that can handle the varying circumstances under which query images are taken. A user might take a picture of a building in which he is interested during any time of the day, under different weather conditions, from any viewpoint. The approach chosen will therefore be required to work under different lighting conditions, be able to handle low resolution images and preferably also be invariant under affine transformations. A solution that meets both the performance criteria and these practical constraints will be presented.

An outline of our approach, which uses local invariant regions, will be discussed, along with a simple baseline classifier providing a reference point for 'naive expectation' performance. The performance of these techniques was evaluated on a personal computer on a standard industrial dataset; the Zurich Building Database (ZuBuD) [10], the results of which will be discussed. On the basis of these results it will be argued that the ZuBuD, although widely used within the literature, is perhaps too easy to serve as a performance reference for our envisaged application. A custom database which does not suffer from this drawback will therefore be introduced for the evaluation of our method on the mobile phone. It will be shown that our method can compete with more computationally elaborate approaches from the literature in terms of classification performance, whilst still running in acceptable time on the low capacity processor of a mobile telephone and requiring only a minimal amount of storage.

2 Local Invariant Regions

In order to identify a building, the intrinsic properties of the object, such as shape, color and texture, have to be compared with those of known buildings. The main difficulty is to get rid of the extrinsic properties, such as scale, viewpoint and illumination conditions.

There are two classes of approaches to object recognition, based on either global or local features. The global approach characterizes the object by its image as a whole. A buildings color distribution is such a global feature. Because many buildings have similar colors it is very unlikely that only color information is sufficient for the task at hand. More sophisticated ways of characterizing objects globally have several problems. In general they are not robust to occlusion, which is unavoidable with objects the size of buildings, not invariant to the viewpoint and they might require a prior segmentation of the image, which is a hard task in the case of buildings.

The other approach characterizes an object by a representations of a number of local features. Usually the features found in an image are compared with the features seen in the database of known objects, where some sort of voting scheme determines which known object fits best. This method is frequently used successfully in literature [2,3,4,6,8,9,11,13]. The advantages of this approach are that it is robust to occlusion and clutter. Local features are more easily described invariantly of scale, viewpoint, and other extrinsic properties.

In the case of building recognition specifically local features prove successful [11,4,2,3,8,5,6]. Recently a more efficient version of SIFT [4] has been proposed, called i-SIFT, for the recognition of buildings on a mobile telephone using a client server architecture [2]. i-SIFT reduces the runtime of the SIFT algorithm by selecting only informative features and reducing the size of the representation of them. Both SIFT and i-SIFT are very robust approaches and i-SIFT in particular has been shown to yield good performance for building recognition[2]. Unfortunately both approaches are less suited for local execution on a mobile phone, since they require numerous Gaussian convolutions. These proved to be a serious execution time bottleneck on the low-end processor available in the average mobile device. Since we explicitly want to avoid the need for client-server communication during the classification process, a method that is more suitable for execution on the mobile phone is required.

Here we will present a novel object recognition method, based on local invariant features, that is optimized for mobile building identification. Our algorithm follows the basic scheme of finding interest points, representing them invariantly and using a voting scheme based on a distance measure between feature representations to determine the best match. The method makes uses of the characteristics of the problem limit the required resources to a minimum, while still performing very well compared to computationally more expensive approaches.

The problem can be described in more detail as follows: Given a low resolution picture of a building taken with a mobile phone, decide which known building is most similar regardless of viewpoint, scale and illumination conditions. From

each known building a set of pictures from various viewpoints is available to create the application. The classification should be done locally on the mobile phone itself within acceptable time. The resolution of the query image is 160×120 , since this is the only available resolution Java software was able to capture on the mobile phone we used. Although this severely reduces the amount of information available for classification, it also reduces the amount of computation and memory needed.

There are some assumptions we can make to make that make problem easier. First of all we can assume that query pictures are always taken upright, which is fair since buildings newer appear rotated. Secondly, pictures are taken approximately under the same vertical angle as the training images. Furthermore, the method exploits the fact that buildings often exhibit many repetitions and planar surfaces.

2.1 Feature Detection

Recognition of objects based on local features requires a method to select interest points that are repeated in different images of the same object independently. The more points that are repeated throughout different images of the object the better, because only regions that are found in both training and query images facilitate correct classification. Regions that only occur in either training or query images can only lead to false associations. Unfortunately it is inevitable that a method detects false interest points, that are not repeated. Extrema in the intensity image have proved to work well as interest points [13] and are very cheap to detect using an optimized non-maximum suppression algorithm. Intensity extrema that also occur versions of the image that are blurred with a Gaussian kernel are more reliable, because they are repeated more often [11]. Gaussian blurring is a time consuming operation on the mobile phone we used (several seconds for a single blur), however when a convolution kernel is used in which all coefficients are powers of 2, bit shift operations can be used to create a fast implementation for this specific kernel.

Other types of interest points that can be found in the literature were less suitable, because they were expensive to detect. Extrema in the difference-of-Gaussian scale space [4], for instance require many expensive convolution operations. Corner points are also more expensive, but also have the drawback that they tend to appear very frequently at places where the object is not planar, which makes it more difficult to characterize those local features invariantly.

2.2 Feature Representation

Given a local intensity extremum, the small image region around this interest point is used as the basis of the local feature that is used for recognition. The shape and size of these regions are determined in such a way that they adapt itself to the viewpoint of the object. The border points of the region correspond

roughly to sharp intensity changes while radiating from the local intensity extremum, as described in [13]. Unlike what is done frequently, we do not fit an ellipse to these points. In small regions where the object is planar, the (perspective) projection of this region is approximately affine. Because we have assumed that the images are not rotated and the training images are taken under the same vertical angle, vertical lines in the region also appear as vertical lines in the training images. The remaining differences can be compensated with a vertical shear and both vertical and horizontal rescaling. A parallelogram of which the left and right sides are held vertically that is fitted to the border points of the region, would therefore capture the transformations that effect the appearance of the region. To make the feature more distinctive we double the size of the fitted parallelogram. This approach is in particular well suited for man-made structures.

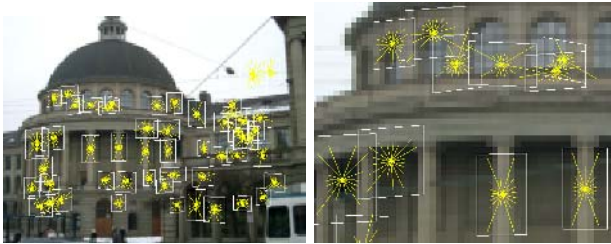


Fig. 1. Examples of found local invariant regions for a building taken from ZuBuD

Unlike others who use ‘Generalized Color Moments’ introduced by Mindru [7] we represent the regions in a more direct way by transforming the image contents in each region to a fixed size square of size 10×10 . The RGB color values of the pixels in this square characterize the region. To make the representation invariant to illumination intensity we divide each value by the sum of the intensities of all pixels in the region. The chosen representation is not rotationally invariant, because this is not needed for our application, making the representation more descriptive than Generalized Color Moments.

In order to reduce the amount of information needed to represent each region, we use Principal Component Analysis to compress the data. Once we have collected all regions from all buildings from the database and represent the pixel data in a normalized region as a vector, we can use PCA to determine a more efficient basis for this feature space. By selecting the first n principal components we can reduce the space required represent a regions significantly and also reduce the amount of computation required for region comparison. We have used the first 30 components, leading to a representation with a space reduction to around 9% while retaining 96% of the original variance of the feature space. Theoretically a drop in performance could be expected in case discriminating features got lost due to this less accurate representation, but this does not appear to be the case in practice.

2.3 Reduction of the Number of Regions

Because we have several images of each building, features which are not repeated throughout these images are likely to be noise or at least not characteristic. Discarding such features therefore not only leads to less storage requirements and faster classifications it also gives a higher accuracy. Repeated regions are not only found because there are more views of each building, but also because many buildings have many repetitions in themselves (i.e. a row of identical windows). It would be a waste of space to store all copies of such repeating regions, so only one prototype is stored. This is achieved by clustering the features found for all images of a building. Singleton clusters are removed, because such features are not characteristic or noise. Of all other clusters, only the centroids are kept as prototypes of the cluster members. We used linkage clustering with an empirically determined threshold based on the maximal distance between the instances.

This novel addition reduces the storage needed to around 20% of the original size, and makes our approach especially suitable for implementation on mobile phones. The clustering step also boosts performance, since it automatically filters out noise that could otherwise reduce classification accuracy.

2.4 Image Classification

When a new image is captured it will undergo the same process of creating a list normalized pixel data of fixed size regions. The data is then projected on the principal component basis. Now the new image can be compared to the database. Classification of the image is performed using a weighted majority voting scheme, wherein each region found within the query image votes for the building belonging to its nearest neighbor in the principal component space.¹ The weight of the vote equals $1/(d + \epsilon)$, where d is the Euclidean distance to the nearest neighbor and ϵ a small constant number to prevent the vote to have infinite weight. Any monotonically decreasing function would work, however experiments have shown that this function gave the highest performance. The query building will be classified as the building with the highest total vote.

3 Baseline Comparison

In order to obtain a basic ‘naive expectation’ measure to indicate the relative performance of our approach, a simple classifier based on normalized RGB (*rgb*) histograms was implemented [12].

For this classifier a histogram is built for the *r* and *g* channel of every building in the database with 100 bins for each channel.² The histogram is then normalized and stored in the database.

¹ Determining the nearest neighbor can be done in $\mathcal{O}(\log n)[1]$.

² Note that the *b* channel contains no extra information since $r + g + b = 1$ and can therefore be dropped.

To classify a new image a histogram is constructed of the query image according to the same parameters as the histograms in the database. The χ^2 distance between the histogram of the query image and every histogram in the database is calculated, after which the query image is classified as being the building for which the χ^2 distance is the smallest.

4 Experiments and Results

To test the viability of our local invariant region approach it has first been implemented in MATLAB, where it was tested on the ‘Zürich Building Database’, ZuBuD[10]. The obtained results will be discussed below. The performance of our approach will be compared to that of other approaches from the literature.

4.1 The ZuBuD Database

The ZuBuD consists of pictures of 201 different buildings taken in the city of Zürich, Switzerland. For every building there are five different views within the database, each of which are 640×480 pixels in size. Differences between the views include the angle at which the picture is taken, relatively small scaling effects and occlusions. Examples of images from this dataset can be found in figure 2. The ZuBuD comes with a standardized query set, consisting of 115 images of buildings occurring in the database. These query images have a resolution of 320×240 and are taken with a different camera under different conditions.

Results on ZuBuD. Since on the mobile telephone we only make use of query images with a low 160×120 resolution, we downsampled the ZuBuD images to this resolution as well. We used these images to test our method. For a fair comparison, we also used these downsampled images for the color histogram approach. The results obtained can be found in Table 1, along with performance of other methods found within the literature. Several things can be noted about these results. First of all the local invariant regions classifier we propose performs well on the ZuBuD. The performance of 91 % is identical to that of the i-SIFT algorithm [2] and of the same order of magnitude as that of most other algorithms mentioned in the literature, despite the fact that it uses very little resources, because it has to run within acceptable time on a mobile phone. Secondly it is interesting to note that the *rgb* histogram based classifier shows a surprisingly high performance on ZuBuD. So high in fact, that it outperforms most methods found in the literature, including our own. Note that this is achieved with only information present in the downsampled images. Intuitively, you suspect that a method based on global color distributions is very sensitive to illumination conditions, scale, occlusions and differences between cameras. Using normalized RGB, the method is invariant to illumination intensity, but the representation is still influenced by the illumination color, which is determined by, amongst others, the weather and



Fig. 2. Examples of different views of a single building from the Zürich Building Database

time of the day. Scale influences the building's representation because the ratio between background colors and building colors is different for a different scale. Occlusions add noise to the representation. When there are different buildings with similar colors, when they are made from the same materials for instance, the discriminative power of color histograms can be expected to be too small to cope with these factors, so the method would fail easily. The high performance of the color distribution approach on ZuBuD can be explained from the fact that the ZuBuD query images are very similar to the reference images in terms of weather, viewing direction and scale. It also helps that many of the buildings in the database are painted in nice pastel colors, which increases the discriminative power of color.

Our own method incorporates shape and texture in its representation of a building, but also relies on color. Illumination intensity is compensated by normalizing the total intensity in a region, but the method is not insensitive to other photometric conditions. In light of the above observations on ZuBuD, we created a custom building database ourselves, which is more realistic for our application. This allowed us to evaluate our method under harder conditions and verify our intuition that global color distributions are not discriminative enough in general. The database also allowed us to test our method on the mobile telephone without traveling to Zürich.

4.2 Custom 'Roeterseiland' Database

The 'Roeterseiland' database consists of images of 7 buildings of the Roeterseiland complex of the University of Amsterdam. From each building between 4 to 11 photos are included, depending on the amount of visible sides of the building and the diversity between the sides. The images have a resolution of 160×120 pixels and are resized from the originals are shot with a 5.0 megapixel camera. The set of query images consists of 45 images, which are taken independently by someone else using the built-in cameras of a Sony Ericsson K700i

Table 1. Performance of the local invariant regions approach on ZuBud, together with results found in the literature

method	performance (% correct)
HPAT indexing [11]	86 %
SIFT [4]	86 %
I-SIFT [2]	91 %
Local Invariant Regions	91 %
Baseline matching [3]	92 %
Sublinear indexing [8]	93 %
rgb histograms	94 %
Random subwindows [5]	96 %
LAF [6]	100 %

**Fig. 3.** Examples of images from the ‘Roeterseiland’ database (left) and query set (right), showing some differences in viewpoints

and a Nokia 6630 mobile telephones. The query images show the same buildings from different angles, at different scales and with various kinds of weather. In several images the buildings are partially occluded by people, cars or trees. A few examples of the images in the ‘Roeterseiland’ database are shown in Fig. 3.

Results on the ‘Roeterseiland’ Database on the Mobile Telephone.

The results obtained on the custom ‘Roeterseiland’ database can be found in Table 2. We begin to note that on this dataset *rgb* does not perform very well, which is more along the lines of expectation than its performance on ZuBuD, reflecting the more realistic quality of the query sets used on our database. Our local invariant region approach still performs quite well, showing 80 % accuracy on the database. This indicates that the performance of our method can not just be ascribed to the lack of difference between training and query images from the ZuBuD, but proves to be reasonably robust. The representation of buildings in our method is invariant to illumination intensity, but not to other photometric properties. The results show that the used features have enough discriminative power to overcome this shortcoming. It might help to use training images taken under identical conditions, to avoid a bias for buildings captured with the same weather for instance.

Table 2. Performance on the ‘Roeterseiland’ database

method	performance (% correct)
Local invariant regions	80 %
<i>rgb</i> histogram	24 %

The algorithm takes less than 5 seconds to classify a building on a Sony Ericsson K700i, and requires only 63 KB bytes of storage for the database (less than 10 KB for a single building). These statistics show that our approach can be efficiently implemented on a mobile phone and that an application performing building recognition locally on the mobile device is indeed feasible in practice.

The number of buildings in the ‘Roeterseiland’ database is small, which influences the performance measured positively. The ZuBuD shows that the method’s accuracy scales to a large number of buildings. The execution time of the method consists largely of the constant time it takes to extract the image’s features. The comparison of each feature with the database is logarithmic in the number of stored regions, when implemented efficiently [1]. In terms of time performance the method can be considered scalable too. Furthermore, in many cases the number of candidate buildings might not be very large. This is the case when the approximate location of the mobile phone is known, using information about the mobile phone network cell for instance.

5 Conclusions

A new local invariant region algorithm for building classification was proposed that could combine robustness with fast performance on a mobile phone while requiring only limited storage. The algorithm was evaluated on a personal computer on the ZuBuD and was shown to be capable of obtaining state of the art results on this database (91 % accuracy).

In order to obtain a baseline performance score as reference material, a very simple classifier using only *rgb* histograms was also implemented. It was shown that this very simple method can outperform most of the known methods from the literature on ZuBuD, indicating that the query images of the ZuBuD database are too easy to differentiate between simple methods and robust advanced methods. We therefore created a small database of our own, consisting of seven buildings and several training and query images for each building. When creating this database we used a high quality camera for the creation of the database and two low-quality built-in mobile phone cameras to create the query images from different viewpoints, at different scales, under different illumination conditions and with realistic small occlusions.

Our algorithm was implemented on a mobile phone (Sony Ericsson K700i) and tested on our custom database along with the baseline *rgb* histogram classifier. For the naive color distribution approach performance performed very poorly on this database (24 % accuracy), whereas the local invariant region algorithm

kept performing well (80 % accuracy). This indicates that the custom database is more challenging and more suitable for obtaining an estimate of the level of performance that can be expected in practice.

References

1. Bentley, J. L., Weide, and Yao, A.: Optimal expected time algorithms for closest point problem. *ACM Transactions on Mathematical Software*, Vol. 6, No. 4, 1980, pp. 563-580.
2. Fritz, G., Seifert, C. and Paletta, L.: A Mobile Vision System for Urban Detection with Informative Local Descriptors. *ICVS '06* 4 (2006) p. 30
3. Goedeme, T., Tuytelaars, T., van Gool, L.: Fast Wide Baseline Matching for Visual Navigation. *CVPR'04* 1 (2004)
4. Lowe, G.: Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision*, 2004.
5. Marée, R., Guerts, P., Piater, J. et al.: Decision Trees and Random Subwindows for Object Recognition. *ICML workshop on Machine Learning Techniques for Processing Multimedia Content MLMM '05*
6. Matas, J., Obdržálek, S.: Object Recognition methods Based on Transformation Covariant Features, XII. European Signal Processing Conference 2004.
7. Mindru, F., Moons, T. and van Gool, L.: Recognizing color patterns irrespective of viewpoint and illumination, *IEEE Conference on Computer Vision and Pattern Recognition*, vol. 1, pp. 368-373, 1999.
8. Obdržálek, S. and Matas, J.: Sub-linear indexing for large scale object recognition. *Proceedings of the British Machine Vision Conference* volume 1, pages 1-10, 2005
9. Obdržálek, S. and Matas, J.: Image Retrieval Using Local Compact DCT-based Representation. *DAGM'03, 25th Pattern Recognition Symposium* September 10-12, 2004, Magdeburg, Germany. *Proceedings of the British Machine Vision Conference* volume 1, pages 1-10, 2005
10. Shao, T. S. H. and Gool, L. V.: Zubud-zurich buildings database for image based recognition. *Technique report No. 260*, Swiss Federal Institute of Technology, 2003.
11. Shao, H., Svoboda, T., Tuytelaars, T. and van Gool, L.: indexing for fast object/scene recognition based on local appearance. *Image and Video Retrieval, Second International Conference, CIVR 2003*, page 71-80
12. Swain, M. and Ballard D.: Color Indexing, *International Journal of Computer Vision*, Vol. 7, pp. 11-32, 1991
13. Tuytelaars, T. and van Gool, L. J.: Wide baseline stereo based on local affinity invariant regions. *British Machine Vision Conference*, 2000.
14. Zhang, W. and Kosecka, J.: Localization based on Building Recognition. *Workshop on Applications for Visually Impaired*, *IEEE Conference, CVPR*, 2005

Adaptive Learning Procedure for a Network of Spiking Neurons and Visual Pattern Recognition

Simei Gomes Wysoski, Lubica Benuskova, and Nikola Kasabov

Knowledge Engineering and Discovery Research Institute,
Auckland University of Technology, 581-585 Great South Rd,
Auckland, New Zealand
{swysoski, lbenusko, nkasabov}@aut.ac.nz
<http://www.kedri.info>

Abstract. This paper presents a novel on-line learning procedure to be used in biologically realistic networks of integrate-and-fire neurons. The on-line adaptation is based on synaptic plasticity and changes in the network structure. Event driven computation optimizes processing speed in order to simulate networks with large number of neurons. The learning method is demonstrated on a visual recognition task and can be expanded to other data types. Preliminary experiments on face image data show the same performance as the optimized off-line method and promising generalization properties.

1 Introduction

The human brain has been modelled in numerous ways, but these models are far from reaching comparable performance. These models are still not as general and accurate as the human brain despite that outstanding performances have been reported [1] [2] [3]. Of particular interest to this research are the models for visual pattern recognition. Visual pattern recognition models can be divided in two groups according to the connectionist technique applied. Most of the works deal with the visual pattern recognition using neural networks comprised of linear/non-linear processing elements based on the neural rate-based code [4] [5]. Here we refer to these methods as traditional methods. In another direction, a visual pattern recognition system can be constructed through the use of brain-like neural networks.

Brain-like neural networks are networks that have a closer association with what is known about the way brains process information. The definition of brain-like networks is intrinsically associated with the computation of neuronal units that use pulses. The use of pulses brings together the definitions of time varying postsynaptic potential (*PSP*), firing threshold (ϑ), and spike latencies (Δ), as depicted in Figure 1 [6]. Brain-like neural networks, despite being more biologically accurate, have been considered too complex and cumbersome for modeling the proposed task. Table 1 shows a general classification of neural models according to the biological accuracy. However recent discoveries on the information processing capabilities of the brain and technical advances related to massive parallel processing, are bringing back the idea of using biologically realistic networks for pattern recognition. A recent pioneer-

ing work has shown that the primate (including human) visual system can analyze complex natural scenes in only about 100-150 ms [7]. This time period for information processing is very impressive considering that billions of neurons are involved. This theory suggests that probably neurons, exchanging only one or few spikes, are able to form assemblies, and process information. As an output of this work, the authors proposed a multi-layer feed-forward network (SpikeNet) of integrate-and-fire neurons that can successfully track and recognize faces in real time [7].

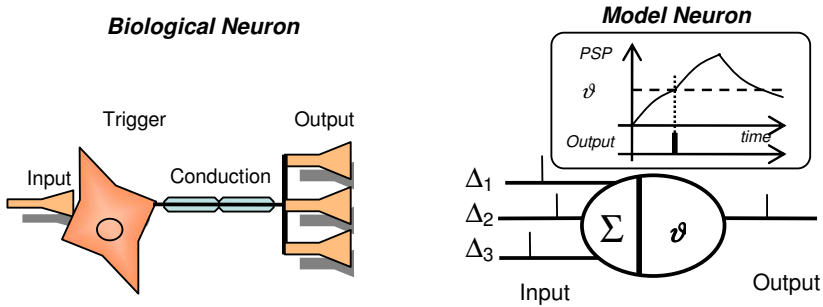


Fig. 1. On the left: Representation of biological neuron. On the right: Basic artificial unit (spiking neuron).

This paper intends to further evaluate the network model SpikeNet proposed in [8] and extend its applicability to perform on-line learning and evolving its functions as streams of information are received by the input nodes. In the next sections the spiking neural network model will be presented and the new learning procedure will be described. The new learning method is applied to the face recognition task and the results are compared with previous work. Discussions and additional required analysis conclude the paper.

Table 1. Classification of artificial neural network models according to the biological relevance

	Biologically motivated	Moderate biological relevance	No biological relevance
Model	Hodgkin-Huxley Multi-compartment Cable theory	Spike Response Model Integrate-and-Fire neuron Izhikevich simple neuron SpikeNet	McCulloch-Pitts Adaline Perceptron
Usage	Tools for neuroscientists	Simulation of large networks Temporal properties and synchronicity of spiking neurons Pattern recognition	Pattern recognition Engineering problems

2 Spiking Network Model

In this section we describe the steps of the biologically realistic model used in this work to perform on-line visual pattern recognition. The system has been implemented

based on the SpikeNet introduced in [7] [8] [9] [10]. The neural network is composed of 3 layers of integrate-and-fire neurons. The neurons have a latency of firing that depends upon the order of spikes received. Each neuron acts as a coincidence detection unit, where the postsynaptic potential for neuron i at a time t is calculated as:

$$PSP(i, t) = \sum \text{mod}^{order(j)} w_{j,i} \quad (1)$$

where $\text{mod} \in (0,1)$ is the modulation factor, j is the index for the incoming connection and $w_{j,i}$ is the corresponding synaptic weight. See [7] [9] for more details.

Each layer is composed of neurons that are grouped in two-dimensional grids forming neuronal maps. Connections between layers are purely feed-forward and each neuron can spike at most once on spikes arrival in the input synapses. The first layer cells represent the ON and OFF cells of retina, basically enhancing the high contrast parts of a given image (high pass filter). The output values of the first layer are encoded to pulses in the time domain. High output values of the first layer are encoded as pulses with short time delays while long delays are given to low output values. This technique is called Rank Order Coding [10] and basically prioritizes the pixels with high contrast that consequently are processed first and have a higher impact on neurons' PSP.

Second layer is composed of eight orientation maps, each one selective to a different direction (0° , 45° , 90° , 135° , 180° , 225° , 270° , and 315°). It is important to notice that in the first two layers there is no learning, in such a way that the structure can be considered simply passive filters and time domain encoders (layers 1 and 2). The theory of contrast cells and direction selective cells was first reported by Hubel and Wiesel [11]. In their experiments they were able to distinguish some types of cells that have different neurobiological responses according to the pattern of light stimulus.

The third layer is where the learning takes place and where the main contribution of this work is presented. Maps in the third layer are to be trained to represent classes of inputs. See Figure 2 for the complete network architecture. In [7], the network has a fixed structure and the learning is done off-line using the rule:

$$\Delta w_{j,i} = \frac{\text{mod}^{order(a_j)}}{N} \quad (2)$$

where $w_{j,i}$ is the weight between neuron j of the 2nd layer and neuron i of the 3rd layer, $\text{mod} \in (0,1)$ is the modulation factor, $order(a_j)$ is the order of arrival of spike from neuron j to neuron i , and N is the number of samples used for training a given class.

In this rule, there are two points to be highlighted: a) the number of samples to be trained needs to be known *a priori*; and b) after training, a map of a class will be selective to the average pattern.

There are also inhibitory connections among neuronal maps in the third layer, so that when a neuron fires in a certain map, other maps receive inhibitory pulses in an area centred in the same spatial position. An input pattern belongs to a certain class if a neuron in the corresponding neuronal map spikes first.

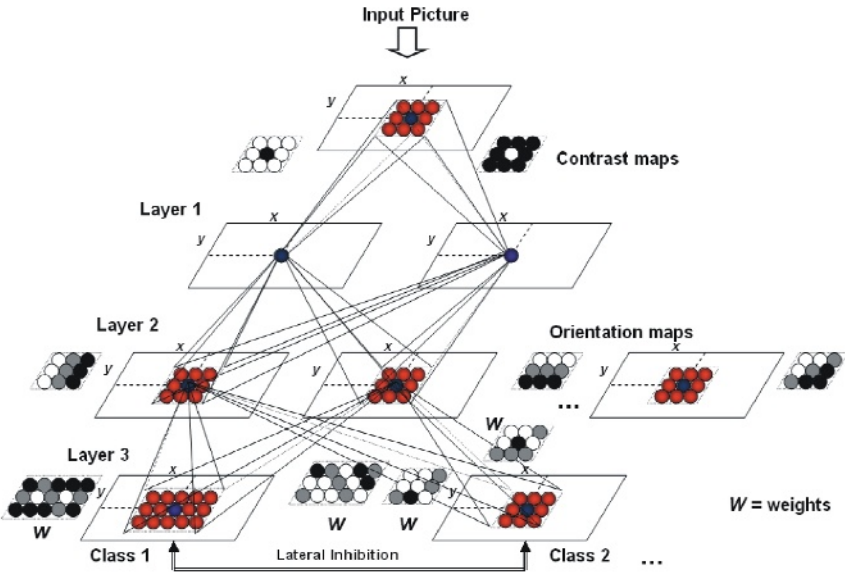


Fig. 2. Adaptive spiking neural network (aSNN) architecture for visual pattern recognition

One of the properties of this system is the low activity of the neurons. It means that the system has a large number of neurons, but only few take active part during the retrieval process. In this sense, through the event driven approach the computational performance can be optimized [8] [12]. Additionally, in most cases the processing can be interrupted before the entire simulation is completed. Once a single neuron of the output layer reaches the threshold to emit a spike the simulation can be finished. The event driven approach and the early simulation interruption make this method suitable for implementations in real time.

3 On-Line Learning and Structural Adaptation

3.1 General Description

Our new approach for learning with structural adaptation aims to give more flexibility to the system in a scenario where the number of classes and/or class instances is not known at the time the training starts. Thus, the output neuronal maps need to be created, updated or even deleted on-line, as the learning occurs. In [13] a framework to deal with adaptive problems is proposed and several methods and procedures describing adaptive systems are presented.

To implement such a system the learning rule needs to be independent of the total number of samples since the number of samples is not known when the learning starts. Thus, in the next section we propose to use a modified equation to update the weights based on the average of the incoming patterns. It is important to notice that, similarly to the batch learning implementation of Equation 2, the outcome is the average pattern. However, the new equation calculates the average dynamically as the input patterns arrive.

There is a classical drawback to learning methods when, after training, the system responds optimally to the average pattern of the training samples. The average does not provide a good representation of a class in cases where patterns have high variance (see Figure 3). A traditional way to attenuate the problem is the *divide-and-conquer* procedure. We implement this procedure through the structural modification of the network during the training stage. More specifically, we integrate into the training algorithm a simple clustering procedure: patterns within a class that comply with a similarity criterion are merged into the same neuronal map. If the similarity criterion is not fulfilled, a new map is generated. The entire training procedure follows 4 steps described in the next section and is summarized in the flowchart of Figure 4.

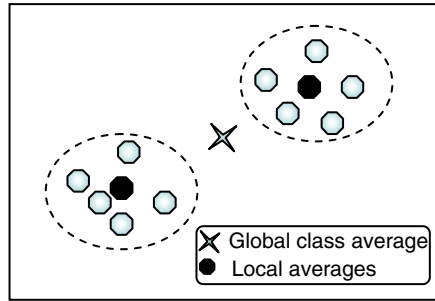


Fig. 3. *Divide and conquer* procedure to deal with high intra class variability of patterns in the hypothetical space of class K . The use of multiple maps that respond optimally to the average of a subset of patterns provides a better representation of the classes.

3.2 Learning Procedure

The new learning procedure can be described in 4 sequential steps:

1. Propagate a sample k of class K for training into the layer 1 (retina) and layer 2 (direction selective cells – DSC);
2. Create a new map $Map_{C(k)}$ in layer 3 for sample k and train the weights using the equation:

$$\Delta w_{j,i} = \text{mod}^{order(a_j)} \tag{3}$$

where $w_{j,i}$ is the weight between neuron j of the layer 2 and neuron i of the layer 3, $\text{mod} \in (0,1)$ is the modulation factor, $order(a_j)$ is the order of arrival of spike from neuron j to neuron i .

The postsynaptic threshold ($PSP_{threshold}$) of the neurons in the map is calculated as a proportion $c \in [0,1]$ of the maximum postsynaptic potential (PSP) created in a neuron of map $Map_{C(k)}$ with the propagation of the training sample into the updated weights, such that:

$$PSP_{threshold} = c \max(PSP) \tag{4}$$

The constant of proportionality c express how similar a pattern needs to be to trigger an output spike. Thus, c is a parameter to be optimized in order to satisfy the requirements in terms of false acceptance rate (FAR) and false rejection rate (FRR).

3. Calculate the similarity between the newly created map $Map_{C(k)}$ and other maps belonging to the same class $Map_{C(K)}$. The similarity is computed as the inverse of the Euclidean distance between weight matrices.
4. If one of the existing maps for class K has similarity greater than a chosen threshold $Th_{simC(K)} > 0$, merge the maps $Map_{C(k)}$ and $Map_{C(Ksimilar)}$ using arithmetic average as expressed in equation

$$W = \frac{W_{Map_{C(k)}} + N_{samples} W_{Map_{C(Ksimilar)}}}{1 + N_{samples}} \tag{5}$$

where matrix W represents the weights of the merged map and $N_{samples}$ denotes the number of samples that have already being used to train the respective map. In similar fashion the $PSP_{threshold}$ is updated:

$$PSP_{threshold} = \frac{PSP_{Map_{C(k)}} + N_{samples} PSP_{Map_{C(Ksimilar)}}}{1 + N_{samples}} \tag{6}$$

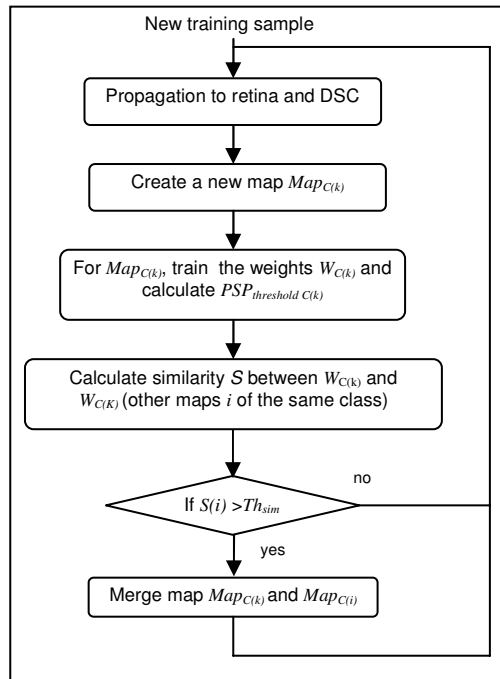


Fig. 4. On-line learning procedure flowchart

4 Experiments and Results

We have implemented the learning procedure proposed in the previous section in a network of spiking neurons as described in section 2. To evaluate the performance and compare with previous work, we used the same dataset as in [7], which is available from [14]. The dataset is composed of 400 faces taken from 40 different people. The frontal views of faces are taken with rotation angles varying in the range of $[-30^\circ, 30^\circ]$.

4.1 Image Preparation

We manually annotated the position of eyes and mouth and used it to centralize the face images. The faces were rotated to align the right and left eyes horizontally. The boundaries of our region of interest (ROI) were then defined as a function of the interocular distance and the distance between the eyes and mouth. The ROI is then normalized to the size 20×30 pixels in greyscale. The 2 dimensional array obtained has been used as input to the SNN. No contrast or illumination manipulation has been performed as previous work demonstrated the good response of the network under the presence of noise and illumination changes [7].

4.2 Spiking Network Parameters

The neuronal maps of retina, DSC and output maps have size of 20×30 . The number of time steps used to encode the output of retina cells to the time domain is set to 100. The threshold for the direction selective cells is set to 600, chosen in such a way that on average only 20% of neurons emits output spikes. The modulation factor $\text{mod} \in (0, 1)$ is set to 0.98. In this way the efficiency of the input of a given neuron is reduced to 50% when 50% of the inputs get a spike. The retina filters are implemented using a 5×5 Gaussian grid and direction selective filters are implemented using Gabor functions in a 7×7 grid. All these parameters were not optimized. Rather, we tried to reproduce as close as possible the scenario described in [7] for comparison purposes.

4.3 Results

Previous work demonstrated the high accuracy of the network to cope with noise, contrast and luminance changes, reaching 100% in the training set (10 samples for each class) and 97.5% when testing the generalization properties [7]. For the generalization experiment the dataset was divided in 8 samples for training and the remaining 2 for test. With the adaptive learning method proposed here, we have obtained similar results for the training set (see Table 2). Varying the similarity threshold Th_{sim} different number of output neuronal maps can be achieved. In the case of similar results, obviously the minimum number of maps is recommended as it saves memory and the processing time is reduced. In this first experiment the postsynaptic threshold $PSP_{threshold}$ is set as $c=0.4 * \max(PSP)$ obtained during the training of a given map. In all tables is presented the best results achieved.

Table 2. Results for the training set according to different similarity thresholds Th_{sim} used to merge maps. All samples (10 images per person, 40 people) were used during training.

Similarity threshold $Th_{sim} (\times 10^{-3})$	0.5	0.714	0.833	1.0
Number of output maps	40	63	213	360
Accuracy (%)	97.00	98.75	99.75	100
False Acceptance Rate (FAR) (%)	0.17	0.12	0.03	0.00
False Rejection Rate (FRR) (%)	1.25	0.00	0.00	0.00

In another experiment, to test the system ability to add on-line output maps for better generalization, we used only 3 sample images from each person for training. The remaining 7 faces views of each person were used for test. Among the dataset faces, we chose manually those samples taken from different angles that appeared to be most dissimilar. Thus, the training set was composed mostly of one face view taken from the left side (30°), one frontal view and one face view taken from the right side (-30°), as depicted in Figure 5. The results are shown in Table 3. In column 2 of Table 3, Th_{sim} is set in such a way that only one output map for each class is created. In such condition, the on-line learning procedure becomes equivalent to the original off-line learning procedure described by Equation 2. Tuning of Th_{sim} for performance, it can be clearly seen the advantage of using more maps to represent classes that contain highly variant samples, as the accuracy of face recognition increases by 6% with a reduction on the FAR.

**Fig. 5.** Example of image samples used for training (30° , frontal and -30°)**Table 3.** Results for the test set according to different similarity thresholds Th_{sim} . Three pictures of each class are used for training and the remaining seven for test.

Similarity threshold $Th_{sim} (\times 10^{-3})$	0.5	0.833	1.0	1.25	2.0
Number of output maps	40	47	80	109	120
Accuracy (%)	74.28	77.49	78.57	80.00	80.00
False Acceptance Rate (FAR) (%)	2.32	2.20	2.18	2.26	1.77
False Rejection Rate (FRR) (%)	0.00	0.00	0.00	0.00	0.00

5 Discussion and Conclusion

We have presented a simple procedure to perform on-line learning in a network of spiking neurons. During learning, new output maps are created and merged based on clustering of intra-class samples. Preliminary experiments have shown that the learning procedure reaches similar levels of performance of the previously presented work, and better performance can be reached in classes where samples have high variability.

As a price, one more parameter needs to be tuned, e.g. Th_{sim} . In addition, more output maps require more storage memory.

With respect to the overall system, the computation with pulses, contrast filters and orientation selective cells finds a close correspondence with traditional ways of image processing such as wavelets and Gabor filters [15] that already have proven to be very robust for feature extraction in visual pattern recognition problems.

In terms of normalization, the rank order codes are intrinsically invariant to changes in contrast and input intensities, basically because the neuronal units compute the order of the incoming spikes and not the latencies itself [7]. Invariance to rotation can be reached with the use of additional neuronal maps, in which each map need to be trained to cover different angles. Here, our learning procedure could be used but, it is important to be aware that it would increase the number of neuronal maps required.

From the biological point of view, while spiking networks can perform all possible operations similarly to traditional neural networks, new ways of connectivity and temporal coding based on biological systems yet to be discovered, can bring new insights to create artificial systems with performance closer to human brains.

As future work we intend to optimize the network parameters and test the learning procedure on more complex datasets. A careful comparison with traditional classification methods and different methods of feature extraction are important to fully understand the potential of the system. In addition, we intend to further extend the network to work in different domains with different types of data.

Acknowledgments

The work has been supported by the NERF grant X0201 funded by FRST (L.B., N.K.) and by the Tertiary Education Commission of New Zealand (S.G.W.).

References

1. Fukushima, K.: Active Vision: Neural Network Models. In Amari, S., Kasabov, N. (eds.): Brain-like Computing and Intelligent Information Systems. Springer-Verlag (1997)
2. Mel, B. W.: SEEMORE: Combining colour, shape, and texture histogramming in a neurally-inspired approach to visual object recognition. *Neural Computation* 9 (1998) 777-804
3. Wiskott, L., Fellous, J. M., Krueger, N., von der Malsburg, C.: Face Recognition by Elastic Bunch Graph Matching: In Jain, L.C. et al. (eds.): Intelligent Biometric Techniques in Fingerprint and Face Recognition. CRC Press (1999) 355-396
4. Haykin, S.: Neural Networks - A Comprehensive Foundation. Prentice Hall (1999)
5. Bishop, C.: Neural Networks for Pattern Recognition. University Press, Oxford New York (2000)
6. Gerstner, W., Kistler, W. M.: Spiking Neuron Models. Cambridge Univ. Press, Cambridge MA (2002)
7. Delorme, A., Thorpe, S.: Face identification using one spike per neuron: resistance to image degradation. *Neural Networks*, Vol. 14. (2001) 795-803
8. Delorme, A., Gautrais, J., van Rullen, R., Thorpe, S.: SpikeNet: a simulator for modeling large networks of integrate and fire neurons. *Neurocomputing*, Vol. 26-27. (1999) 989-996

9. Delorme, A., Perrinet, L., Thorpe, S.: Networks of integrate-and-fire neurons using Rank Order Coding. *Neurocomputing*. (2001) 38-48
10. Thorpe, S., Gauguier, J.: Rank Order Coding. In: Bower, J. (ed.): *Computational Neuroscience: Trends in Research*. Plenum Press, New York (1998)
11. Hubel, D.H., Wiesel, T.N.: Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. *J. Physiol*, 160 (1962) 106-154
12. Mattia, M., del Giudice, P.: Efficient Event-Driven Simulation of Large Networks of Spiking Neurons and Dynamical Synapses. *Neural Computation*, Vol. 12 (10). (2000) 2305-2329
13. Kasabov, N.: *Evolving Connectionist Systems: Methods and Applications in Bioinformatics, Brain Study and Intelligent Machines*. Springer-Verlag (2002)
14. <http://www.cl.cam.ac.uk/Research/DTG/attarchive/facedatabase.html>
15. Sonka, M., Hlavac, V., Boyle, R.: *Image Processing, Analysis, and Machine Vision*, 2nd edn. (1998)

Interactive Learning of Scene Context Extractor Using Combination of Bayesian Network and Logic Network

Keum-Sung Hwang and Sung-Bae Cho

Dept. of Computer Science, Yonsei University,
Shinchon-dong, Seodaemun-ku,
Seoul 120-749, Korea
{yellowg, sbcho}@sclab.yonsei.ac.kr

Abstract. The vision-based scene understanding technique that infers scene-interpreting contexts from real-world vision data has to not only deal with various uncertain environments but also reflect user's requests. Especially, learnability is a hot issue for the system. In this paper, we adopt a probabilistic approach to overcome the uncertainty, and propose an interactive learning method using combination of Bayesian network and logic network to reflect user's requirements in real-time. The logic network works for supporting logical inference of Bayesian network. In the result of some learning experiments using interactive data, we have confirmed that the proposed interactive learning method is useful for scene context reasoning.

1 Introduction

Bayesian network (BN) is a useful tool for modeling causal judgment and inference processes [1], and it receives increasing attention in the vision-based scene understanding area, where it recognizes contexts by reasoning detected objects and features from vision data to understand scene. Bayesian network is also robust to real-world situations because the probabilistic approach manages well uncertain data and supports multiple directional inferences.

Scene understanding is the task of understanding a scene beyond single-object recognition. A scene understanding is determined by constructing a description of the scene in terms of concepts provided in a conceptual knowledge base. It still remains a difficult problem because of complexity and uncertainty of real-world.

In the vision-based scene understanding area, automatic learning is important because not only expert knowledge but also domain data collected in real-world are used for modeling inference model. However, it is not easy since the actual environment of a scene understanding agent has various uncertain data and causes continuous user's requests. In this paper, we propose an interactive learning method to adopt user's requests to inference model by learning Bayesian network with logic network (LN). Because Bayesian network training data and interactive data have different features as follows, we exploit the learning method of logic network. The characteristics of Bayesian network training data:

- Complete data: have values of all nodes.
- Large quantity: require a lot of data for training BN parameters. A node requires 100 data for each distinct instantiation of the parent set for 1% error range.
- Case data: each datum is a case of available states (possible values of node), so the gathered data set reflects probabilistic causalities.

The characteristics of interactive data:

- Incomplete data: have values of a few nodes.
- Small quantity: not easy to collect enough data for training BN parameters
- Logical data: beneficial to design structure.
- Changeable: imply user's opinion, so it is profitable to manage them specially without direct modification of BN parameters.

2 Related Works

Some probabilistic approaches are studied recently to solve the vision-based scene understanding problem as follows;

- T. M. Strat *et al.* (1991): Assumes that a target object is defined by several shape models and can be extracted some local features.
- A. Torralba *et al.* (2003): Recognizes scenes using Hidden Markov model from the visual feature vectors.
- B. Neumann *et al.* (2003): Researches for description logics and framework for high-level scene understanding and interpretation, modeling based on detected objects.
- M. Marengoni *et al.* (2003): Selects visual function sets automatically based on hierarchical BN on aerial picture recognition system - Ascender I.
- J. Luo, A. E. *et al.* (2005): Detects natural objects in outdoor scenes based on probabilistic spatial context[2].

However, the proposed probabilistic model demands an enormous amount of training data or expert's assistance. For this reason, they are difficult to adapt several interactive data effectively added by a user-feedback. There were researches for learning the Bayesian model. Adaptive Bayesian network using revised backward propagation method is researched [3], and Bayesian network refinement method that adapts Bayesian networks using minimum description length score metric is proposed [4].

- B.P.L. Lo *et al.* (2003): Adaptive BN: adapts BN using re-training technique - revised backward propagation[3].
- W. Lam (1998): Bayesian network refinement technique: using MDL score metric that minimizes distance between network structure and data-set[4].

As these methods require complete and sufficient amount of Bayesian network data, and it is not suitable to learn interactive data since they have small quantity and many missing data, and they do not concern the features of interactive data, so we propose a learning method with due consideration to them.

3 Interactive Learning of Bayesian Network Module Using Logic Network

In this section, we propose a BN+LN model (combination of Bayesian network and logic network) to expand or update inference model with collected interactive data. Figure 1 shows the proposed model. In the proposed method, a logic network plays a role of supporting inference on the posterior stage of Bayesian network. The computational result of logic network covers that of Bayesian network.

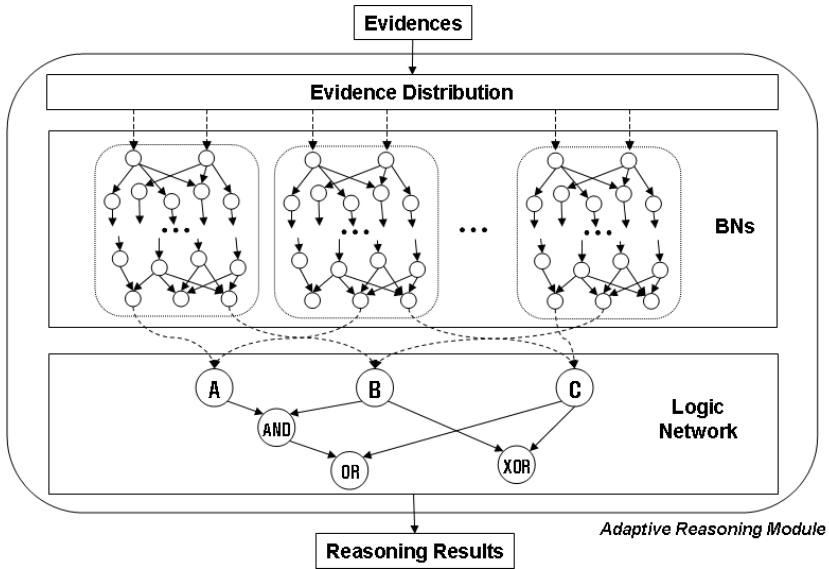


Fig. 1. The BN+LN model

3.1 Bayesian Network

A Bayesian network have a shape of DAG (directed acyclic graph) expressing the relations of nodes and describes a large probabilistic relations with CPTs (conditional probability tables) constrained by the structure. The belief value using the given evidence set E on Bayesian network, $Bel(h)$ is calculated by Bayes' Rule such as a formula (1).

$$Bel(h) = P(h|E) = \frac{P(E|h)P(h)}{P(E)} = \frac{P(E \cap h)}{P(E)}. \tag{1}$$

where h is the hypothesis of a node state. The probability set is computed by a chain Rule such as formula (2).

$$\begin{aligned} P(x_1, x_2, \dots, x_n) &= P(x_1)P(x_2|x_1) \dots \\ &= P(x_1)P(x_2|\pi_2) \dots P(x_n|\pi_n). \end{aligned} \tag{2}$$

where x_i is the i -th node and π_i is the parent of the node i .

3.2 Logic Network

A logic network is a structure for expression of input /outputs or Boolean computations of digital circuit. It has input/output nodes, internal nodes with a logical function and a directed acyclic graph structures that indicates data stream [5]. A node function is composed of multiple inputs and 1 output and a logic function. The internal functions of logic network are listed in Table 1.

Table 1. The internal node function list of logic network. Where v_i means a result value of logic calculation, which is a Boolean type, possible only *true* and *false*. n_i and s_i indicate node and state. *StateEq*(\cdot) and *IsInStateList*(\cdot) work for input value selection.

Function name	Description
<i>NOT</i> (v)	If ($v = false$) then <i>true</i> else <i>false</i>
<i>AND</i> (v_1, v_2, \dots, v_n)	If (every $v_i(1 \leq i \leq n) = true$) then <i>true</i> else <i>false</i>
<i>OR</i> (v_1, v_2, \dots, v_n)	If (any one of $v_i(1 \leq i \leq n) = true$) then <i>true</i> else <i>false</i>
<i>StateEq</i> (n_1)	If (the state of $n_1 = true$ OR the state of $n_1 = yes$) then <i>true</i> else <i>false</i>
<i>StateEq</i> (n_1, s)	If (the state of $n_1 = s$) then <i>true</i> else <i>false</i>
<i>IsInStateList</i> (n_1, n_2)	If (the state of $n_1 = true$) AND (the name of $n_1 \in$ the state list of n_2) then <i>true</i> else <i>false</i>

3.3 Interactive Learning Method

We propose a method that adapts logic network and combine it with Bayesian network to deal with the interactive data collected from the inference module application process. In the proposed method, a logic network works for supporting logical inference of Bayesian network. The detailed operations are given as follows:

- Interaction with user: Interacting with user, collecting new data/feedback from user. In this paper we uses only predefined sentences.
- Reasoning: Inferring contexts that user wants for. If the reasoning is not available the system requires the user to feedback.
- Causality extraction: Defining logical relations of variables. The structure of sentence decides logics by predefined rules.
- Variable extraction: If the variable that is not defined is detected it declares a new variable as evidence-variable or result-variables with its role.
- Adapt logic network: Expressing logical relation of variable as a network and adapt the previous network.
- Update module: Updating the adapted logic network.

4 Experiments

We have applied additional evidences given as interactive data to the proposed interactive learning method of LN+BN model.

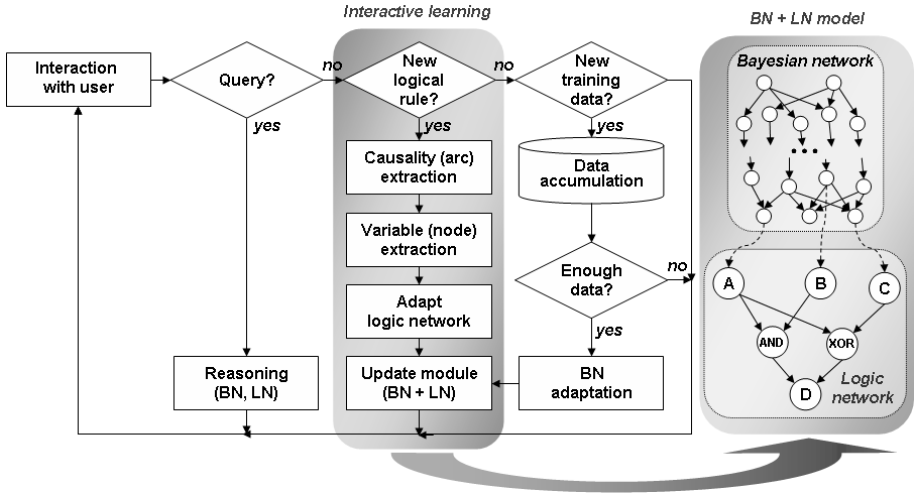


Fig. 2. The proposed interactive learning process

4.1 A Case Study

The first experiment is a case study to observe learning result of the proposed method. We defined a scene domain (11 places and 25 objects) and designed the place-recognition Bayesian network based on detected object (at the left side on figure 3) [6]. Then, we have added 8 logical interactive data for interactive learning of BN+LN model. Table 2 denotes a part of the interactive data used for learning and their extracted logic rules by the proposed interactive learning process. For the smooth experiments we have used predefined sentences and extraction rules. Figure 3 shows an experimental result of adapting given interactive data using the proposed interactive learning method. In the figure, it can be seen that some object nodes and logical inference functions are augmented. In fact, the logic relations of the logic network can be expressed by Bayesian network, but they do not require complex probability values, and a Bayesian network represents inefficiently the node that has many parents, for example "Indoor" node requires $4,096 (= 2^{12})$ CPVs (conditional probability values). The complexity of the number of CPVs is $O(k^N)$ and it is calculated as follows;

$$\prod_{i \in P \cup I} N_i, \tag{3}$$

where P is parents set, I is the self-node, N_k is the number of states of the node k .

4.2 Performance Test

To evaluate the performance of the proposed method, we experimented with interactive data that contain certain agreement/disagreement-evidences for place

Table 2. A part of the 8 interactive data used for learning BN+LN

Object	Interaction Data	Extracted Logic
Elevator	If it is narrow place, has door and no ground objects, and It is linked corridor, it is elevator.	place shape=narrow AND ground object=no AND door=yes AND linked=corridor → elevator. Needs more information about ground object.
Ground object	Air conditioner, garbage can, bookshelf, dresser, chair, lectern, partition, table and castor whiteboard are ground objects.	AirConditioner, garbageCan, bookshelf, dresser, chair, lectern, partition, table, castorWhiteboard → ground-Object
Hall	Hall is linked to corridor and has a door.	NOT (linked=corridor AND door=yes) → NOT hall
Corridor	Corridor is long and linked to an indoor place.	NOT (place shape=long AND linked=indoor) → NOT corridor
Lecture room	Lecture room has a lectern and has a wall whiteboard or a screen.	NOT ((wall whiteboard=yes OR screen=yes) AND lectern=yes) → NOT lectureRoom
Seat place	Seat-place includes chair, bench and sofa.	chair, bench, sofa → seatPlace

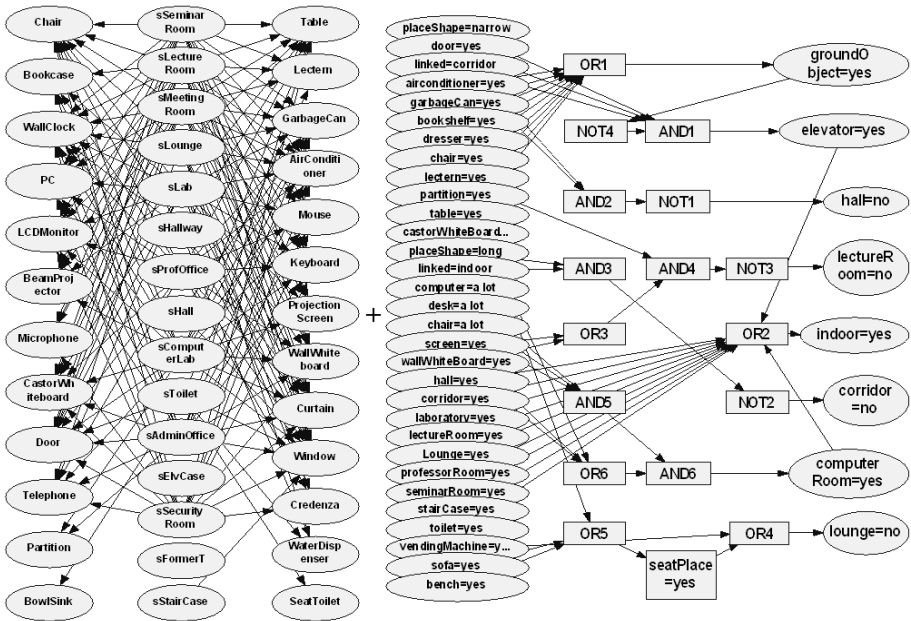


Fig. 3. The designed Bayesian network (left) and the learned logic network (right)

recognition. Where, an agreement-evidence indicates the evidence that is sufficient for an affirmation of fact, and a disagreement-evidence denotes the evidence that is sufficient for negation of a fact. Especially, a negation-logic is usually not used in Bayesian network, because it requires a complex Bayesian network model, while logic network requires a simple model. We experimented each case when k (=evidence size) evidences of 25 objects are discovered. We used totally 3 agreement-evidences and 79 disagreement-evidences for 11 places. When the evidence size is 4 we adopted 2,300 random cases since the number of evidence combination is too large.

Table 3 shows the results, in which the performance of the proposed method is better than the original Bayesian network at all cases except the case of 1-evidence because the interactively-learned logic rules support more accurate inference with relatively smaller evidences.

Table 3. The experimental results with agreement/disagreement-evidences. Abbreviations: Precision = $TP/(TP + FP)$, Recall = $TP/(TP + FN)$, TP =True positive, FP =false positive, FN =false negative, TN =true negative error.

Runs	25 (${}_{25}C_1$)		300 (${}_{25}C_2$)	
Evidence size	1		2	
Method	BN	LN+BN	BN	LN+BN
FP	72%	72%	48%	45%
TN	28%	28%	52%	55%
TP	100%	100%	97%	100%
FN	0%	0%	3%	0%
Precision	58%	58%	67%	69%
Recall	100%	100%	97%	100%
Runs	2,300 (${}_{25}C_3$)		random 2,300	
Evidence size	3		4	
Method	BN	LN+BN	BN	LN+BN
FP	33%	26%	17%	7%
TN	67%	74%	83%	93%
TP	93%	100%	90%	100%
FN	7%	0%	10%	0%
Precision	74%	80%	84%	94%
Recall	93%	100%	90%	100%

5 Concluding Remarks

We propose an interactive learning method using logic network to apply user's requests to Bayesian inference model. The experimental result shows that the proposed interactive learning method is useful for incremental scene context extraction by supporting addition of new context nodes and logical inference rules, so it causes performance improvement. The proposed method might be a good complement for interactive learnable Bayesian network.

In the future work, we would like to apply the proposed method to the more complex and practical applications. We can also compare the method to various Bayesian network adaptation techniques.

Acknowledgement. This research was supported by the Ministry of Information and Communication, Korea under the Information Technology Research Center support program supervised by the Institute of Information Technology Assessment, IITA-2005-(C1090-0501-0019).

References

1. K. B. Korb and A. E. Nicholson, *Bayesian Artificial Intelligence*, Chapman & Hall/CRC, 2003.
2. J. Luo, et al., "A Bayesian network-based framework for semantic image understanding," *Pattern Recognition*, vol. 38, pp. 919-934, 2005.
3. B.P.L. Lo et al., "Adaptive Bayesian networks for video processing," *Int Conf. on Image Processing*, pp.889-892, 2003.
4. Wai Lam, "Bayesian network refinement via machine learning approach," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 20, no. 3, pp. 240-251, March 1998.
5. P.G. Bishop, "Estimating PLC logic program reliability," *Safety-critical Systems Symposium*, Birmingham, 17-19 Feb., 2004.
6. Y.-S. Song, S.-B. Cho, and I.-H. Suh, "Activity-object Bayesian networks for detecting occluded objects in uncertain indoor environment," *Lecture Notes in Computer Science*, vol. 3683, pp. 937-944, 2005.
7. B. Neumann, *A conceptual framework for high-level vision*, Bericht, FB Informatik, FBI-HH-B245/03, 2003.
8. A. Torralba, et al., "Context-based vision system for place and object recognition," *Int. Conf. Computer Vision*, pp. 273-280, 2003.

Adaptative Road Lanes Detection and Classification*

Juan M. Collado, Cristina Hilario, Arturo de la Escalera, and Jose M. Armingol

Intelligent Systems Lab, Universidad Carlos III de Madrid, Spain
{jcollado, chilario, escalera, armingol}@ing.uc3m.es

Abstract. This paper presents a Road Detection and Classification algorithm for Driver Assistance Systems (DAS), which tracks several road lanes and identifies the type of lane boundaries. The algorithm uses an edge filter to extract the longitudinal road markings to which a straight lane model is fitted. Next, the type of right and left lane boundaries (*continuous, broken or merge* line) is identified using a Fourier analysis. Adjacent lanes are searched when broken or merge lines are detected. Although the knowledge of the line type is essential for a robust DAS, it has been seldom considered in previous works. This knowledge helps to guide the search for other lanes, and it is the basis to identify the type of road (*one-way, two-way or freeway*), as well as to tell the difference between allowed and forbidden maneuvers, such as crossing a continuous line.

1 Introduction

The goal of Intelligent Transportation Systems is to increase security, efficiency, and comfort of the transport, by improving the functionality of vehicles and roads using information and communication technologies.

The development of a DAS able to identify dangerous situations involves deep analysis of the environment, including elements such as road, vehicles, pedestrians, traffic signs, etc. and the relationships among them. For instance, detecting a vehicle in the scene represents a risky situation, but the risk is higher when the vehicle is in an adjacent lane in a two-way road – i.e. it is oncoming – than when it is in a freeway. Likewise, there are differences between crossing a broken line in a freeway and crossing a continuous line in a two-way road. However, most current DAS cannot tell the difference between these situations.

Regarding the perceptual system, a DAS may be based on passive sensors like cameras or active sensors such as radar or lidar. The cameras give much more information, but the radars and lidars have better performance in bad weather conditions. However, according to statistics most accidents occur during daylight and with good weather conditions. This fact makes computer vision an adequate perception system in this case.

This paper presents the Road Detection and Classification module of the IvvI project (Intelligent Vehicle based on Visual Information). Its goal is to

* This work is partially supported by the Spanish government through the CICYT project ASISTENTUR.

automatically detect the position, orientation and type of road lanes that the camera can be see. This is achieved by identifying the type of lane boundaries (*continuous*, *broken* or *merge*), and looking for adjacent lanes when a broken or merge line is detected. This perceptual ability pretends to be the basis of a better evaluation of the potential danger of a situation.

1.1 Previous Work

Road detection algorithms for marked roads can be classified in two groups:

1. *Model-based* methods follow a top-down approach. Their main advantage is that the lane can be tracked with a statistical technique, thus, false detections are almost completely avoided. However, as they follow a top-down approach, only the features included in the model are found. Therefore, it is difficult to build a model that is able to adapt to new roads or environment conditions.
2. *Feature-based* methods follow a bottom-up approach. All the features that are in the image are subject to be found, but noise can generate false detections.

Most of the current research effort moves towards adjusting high order models to the lane shape. The goal is to extract accurate information, overcoming the instabilities and noise sensibility typical of more complex models such as the 4D [7] and zero-bank [6] [11]. In [14] horizontal curvature is modeled as a cubic, and the lane is tracked with an enhanced CONDENSATION algorithm [10]. Similarly, in [4] the horizontal curvature of the road shape is also modeled as a third order polynomial, and the vertical curvature as a second order polynomial. Other works try to adjust splines [15] or snakes [17] [16], but these are more difficult to track.

On the other hand, there are few works on longitudinal road markings classification and road type recognition, although this information is essential. Few works consider the existence of other lanes, which is directly related to the road type. The direction of vehicles on other lanes, the possible maneuvers and the speed limit, are just some examples of facts that depend on the road type.

In [3] a six parameter model that merges shape and structure is used. The shape is modeled as a second order polynomial, and the structural model considers the road line as a square waveline, with its period, duty cycle and phase. The parameters can be tracked from frame to frame, but the algorithm requires an initialization step that is very time consuming. Besides that, only one lane boundary mark is fitted to each frame. In [13] road lines are roughly classified in *solid* or *broken*, by analyzing the gaps between the measurement points. If the gap overcomes a threshold the road marking is classified as broken. Thus, the algorithm can easily be confused with any obstacle or structured noise that occludes the marking line, such as shadows or other vehicles. This work also tries to estimate the left and right adjacent lanes assuming that some of their parameters are identical to those of the central lane. Likewise, in [1] an array of probabilities which defines the presence of lateral lanes is kept. The lanes

are numbered, and another array stores the identification number of the lane in which the vehicle is traveling.

In short, these methods can detect any number of lanes, but there is a need of an external technique that indicates to the algorithm how many lanes to search for, and where they can be located (right or left). The difficulty arises from the use of a top-down approach without considering the lane marking type.

2 Tracking and Adaptative Detection of Road Lanes

Figure 1 shows the flow chart of the algorithm proposed in this paper. In brief, this algorithm goes through the following steps. First, it generates a bird-eye view of the road through a perspective transformation. Second, it segments the pixels which belong to longitudinal road markings. Next, the right and left boundaries of the ego-lane are extracted by the Hough Transform [9]. And finally, the pitch angle is corrected, and the lane boundaries are classified in *continuous*, *broken*, and *merge*. If a lane border is identified as a *broken* or *merge* line, the algorithm keeps searching for other lane boundaries until a continuous line is found or the image boundary is reached. These steps are explained in depth in the following sections.

2.1 Perspective Transformation

The image analysis can be done in two different reference systems. Specifically, the road can be analyzed from the *car view* image (Fig. 2(a)), as in [12], or from a *bird-eye view* after a perspective transformation [2], assuming that the world is flat (Fig. 2(b)).

The bird-eye view is easier to process because road lines appear parallel, have constant width, and are mainly vertical. Besides, every pixel of the image appears in world coordinates. This is very useful for the road lines classification, as will

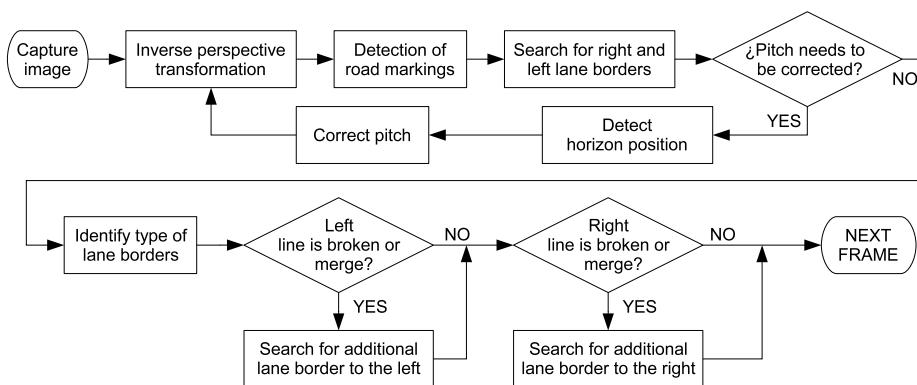


Fig. 1. Flow chart of the proposed algorithm

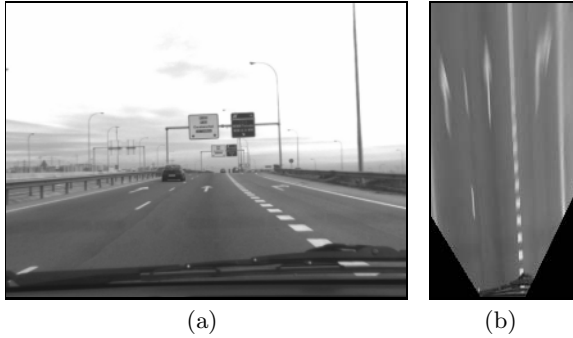


Fig. 2. (a) On-board camera-view; (b) Bird-eye view

be explained in section 2.5, and this is the main reason why we have chosen this reference system. Furthermore, the size of the bird-eye (256×128) is much smaller than the original image (640×480), so that its processing is considerably faster.

However, the bird-eye view presents calibration problems. If the extrinsic calibration parameters of the vision system – i.e. its position and orientation in world coordinates – are not well calculated, the flat road assumption is violated, and the bird-eye view image will show converging or diverging lines instead of parallel ones. This leads to a bad calculation of the lane position and the lane orientation. In order to overcome these problems, an auto-calibration algorithm based on evolutionary techniques is used [5]. This algorithm gives a first estimation of the extrinsic parameters of the vision system, and is run when the cameras are installed in the vehicle. Thereafter, the pitch angle is corrected in every frame by detecting the height of the horizon, as explained in section 2.6.

2.2 Road Model

The road model comprises two parts, the road geometry (linear, parabolic, etc.) and the road type (one-way, two-way or freeway with a variable number of lanes). As has been said in Sect. 1.1, many geometric road models have been extensively researched, but there is little emphasis in road type interpretation. This algorithm is designed to automatically classify the road lines, detect the number of lanes, and track them. Thus, the main contribution of this paper is the automatic road type detection. At present, the algorithm works in freeways with a variable number of lanes.

With regard to road geometry, in this paper we consider the road to be straight for three main reasons. First, straight lines are faster to detect and faster to track than higher order models. They can be robustly and quickly extracted with the Hough Transform, a technique that can hardly be applied to more complex models in real time. Second, it eases other processes such as auto-calibration, the tracking of the pitch angle, and, above all, the road lines classification. Finally, it is a reasonable approximation in the nearby region of the road.

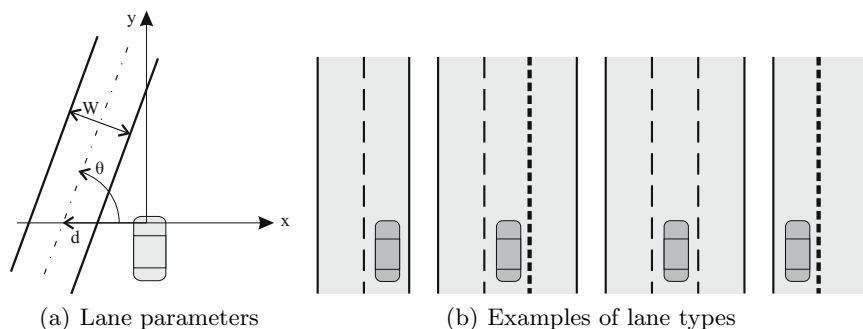


Fig. 3. Road model

The geometric road model is shown in Fig. 3. It has three parameters: d is the distance to the center of the ego-lane, θ is the yaw of the vehicle with regard to the lane, and W is the lane width. The road type model considers the road as a freeway with up to three lanes, in which the lane boundaries can be one of three types: *continuous*, *broken* or *merge*.

2.3 Road Markings Detection

This step extracts from the original image the pixels that are candidates to belong to a road line. Road lines can be considered as bright bands over a darker background. As the lane curvature is small in the nearby region of the road, these lines are mainly vertical in the bird-eye view image of the road. Therefore, the search for pixels that belong to road markings consists of looking for dark-bright-dark transitions in the horizontal direction.

The borders of the image are extracted with a spatial filter based on the ideas of the Canny border extractor, which offers a good signal-noise ratio, compared to other border extractors. This filter uses the intermediate steps of the Canny filter to estimate the orientation of the border, and is used to obtain a horizontal gradient image. Thus, the borders that are not essentially vertical are discarded.

Figure 4 shows how road markings produce two opposite peaks within a certain range of distances in a row of the gradient image. The algorithm scans the horizontal gradient image row by row, searching for a pattern composed of a pair of peaks of opposite sign which are spaced a distance equal to the line width. The line width is considered to be between ten and fifty centimeters in world coordinates. When this pattern is found, the middle point is labeled as a road marking.

2.4 Adaptative Road Lanes Detection

Next, the Hough Transform is used to detect straight lines. Compared to other model fitting methods, the Hough Transform is very robust as it uses global

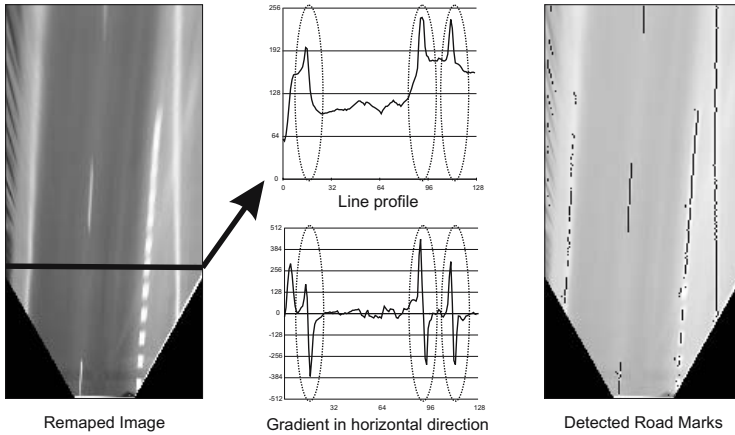


Fig. 4. Detection of pixels belonging to road markings

information. Then, it can easily detect the road lines even though they are broken or partially occluded. In addition, when the model is simple and the image is small, it is fast enough to be applied in real time.

The usual ρ - θ parameterisation is used for straight lines. As the lines are mainly vertical, in the accumulator matrix the parameter θ is constrained to the range $[-15^\circ, +15^\circ]$. Once the accumulator is calculated, only some regions of interest (ROIs) are scanned for local maximums. The ROIs are delimited with the predictions of the Kalman filter. This fact speeds up computation, and avoids interferences with other features outside of the search region.

Kalman filter is used to track five variables: the lateral position (d) and speed (\dot{d}) of the vehicle with regard to the center of the ego-lane, the orientation (θ) of the vehicle respect to the lane, the angular speed ($\dot{\theta}$), and the lane width (W). The width and height of the ROIs – i.e. the interval in ρ and θ – are calculated from the confidence interval of the lateral position and the orientation of the vehicle, respectively.

For the first frame, only two ROIs are considered. These ROIs are big enough to contain the right and left boundaries of the ego-lane. If several lines are found in the same ROI, the algorithm tries to match each line or one ROI with a line of the other ROI, i.e., find the opposite lane border, which should have the same orientation. The best match is used as the initial observation that will be tracked. If no lines can be matched, the most voted line is used. For subsequent frames, if several lines are found, Kalman filter will select the observation that gives the best χ^2 -test result.

Once the ego-lane has been detected, its lane boundaries are classified in continuous, broken or merge (as explained in Sect. 2.5). When a line is identified as broken or merge, an additional ROI is created to look for a new road line that should be at the same side, and separated a distance equal to the lane width (Fig. 5).

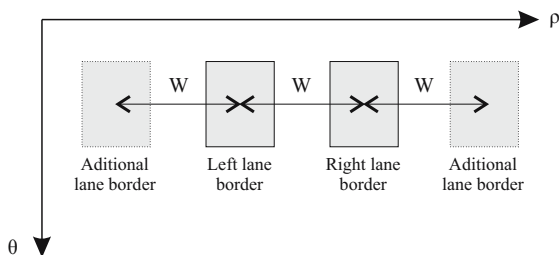


Fig. 5. Regions of interest in the accumulator of the Hough Transform

2.5 Road Lines Classification

The extracted lines are classified in the different types of lines that are found on roads. The main difficulty of this task is the lack of international standardization of the length and frequency of the white stripes in broken lines. However, most roads have three basic line types already mentioned, namely: *continuous*, *broken* and *merge*.

In order to explain this stage of the algorithm, the three lines showed in Fig. 6 will be used as examples. Each of them represent one of the three classes that are being considered.

The intensity line profile for each detected line (right column of Fig. 7) is not a good data to feed the frequency analysis, because its appearance changes substantially with the environment conditions. Besides, the resolution of the bird-eye view in the distance is poor. This effect represents an inconvenience in the merge lines, which appear blurred far ahead and could even look like a continuous one (Fig. 7(a)). Besides, the power spectrum (left column of Fig. 7) presents a tiny peak at the specific frequency of the merge line.

It is more robust to obtain the line profile from the thresholded image given by the road markings detection step (Fig. 6(b)), which is showed on the left side

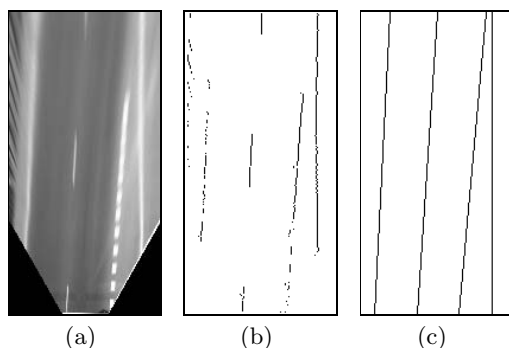


Fig. 6. (a) Remapped image; (b) Detected Road Markings; (c) Lines detected by Hough Transform

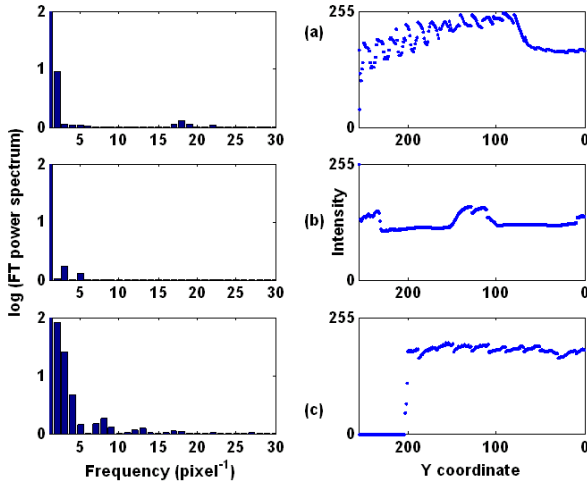


Fig. 7. (right) line profile extracted from the intensity image (Fig. 6(a)); (left) Power spectrum of the Fourier analysis; (a) merge line; (b) broken line; (c) continuous line

of Fig. 8. Again, the right side of the figure shows the power spectrum of Fast Fourier Transform applied to the line profile vector. The results show that a clear sharp peak appears in the Fourier Transform power spectrum when the line is broken, and that the value of the frequency associated to that peak gives the line type (broken or merge). These peaks are showed on the left side of Fig. 8(a) and Fig. 8(b) with arrows pointing at them. No significant peaks are present when the line is continuous (Fig. 8(c)). It can now be seen that the peaks are sharper and much easier to detect. It has been heuristically found that only the first 21 frequencies are significant in this analysis.

Thus, the classification is performed by scanning the first 21 frequencies. Two requirements are needed in order to classify a line as broken or merge:

1. In the first place, a peak must be found within a certain range of frequencies. Two different frequency ranges have been specified. The broken vertical lines on the left side of Fig. 8(a) and Fig. 8(b) show the limits for merge and broken lines, respectively.
2. In second place, the peak must overcome a threshold, which depends on the frequency interval, as the height of the peak decreases as the frequency increases. On the left side of Fig. 8, a horizontal line shows the threshold for each frequency interval. Figure 8(c) shows that no peak exceeds the threshold in neither of the specified frequency ranges when the line is *continuous*.

2.6 Pitch Angle Correction

The extrinsic parameters of the vision system are calculated during installation, but these parameters suffer small drifts during driving, specially the pitch angle

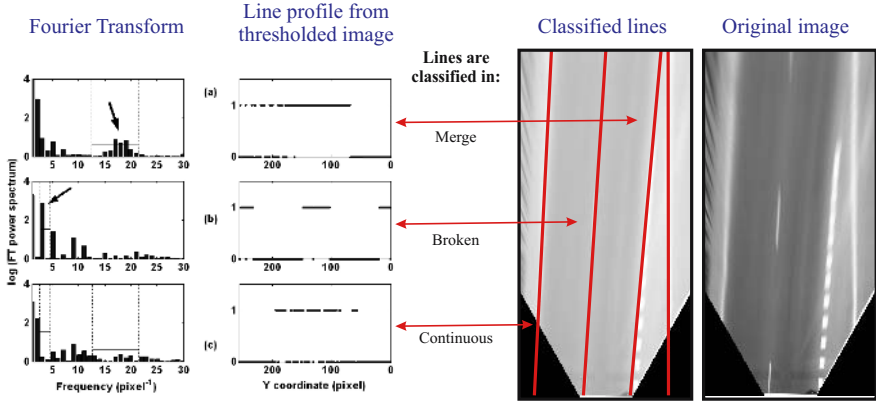


Fig. 8. Fourier analysis for the line profile; (a) merge line; (b) broken line; (c) continuous line

due to the usual swaying of the vehicle, e.g. in sudden braking, dips, etc. This is the most critical parameter because it notably distorts the bird-eye view obtained through the perspective transformation.

In order to correct the pitch angle, the image is processed twice. First, two lane boundaries are detected and its intersection point calculated. This point should belong to the horizon line. These lines do not need to be classified since they are only used to estimate the horizon height. Then the pitch angle (ϕ) is given by the equation:

$$\phi = \arctan \left(\frac{y_{horizon} - y_{center}}{f} \right) \tag{1}$$

$y_{horizon}$ is the y coordinate of the horizon line in pixels,
 where: y_{center} is the y coordinate of the center of the CCD in pixels, and
 f is the focal distance in pixels.

With the updated pitch angle, the bird-eye view is regenerated, now with correct parameters. The corrected image is processed according to the steps explained in previous sections.

3 Results

This algorithm has been tested in the IvvI platform. IvvI (Fig. 9) is a research platform for the implementation of systems based on computer vision, with the goal of building an Advanced Driver Assistance System (ADAS). It equipped with a stereo-vision system composed of two B&N progressive scan cameras used for road, vehicle, and pedestrian detection, a color camera, used for traffic signs detection, a GPS to measure speed, and a processing system composed of two Pentium IV computers.

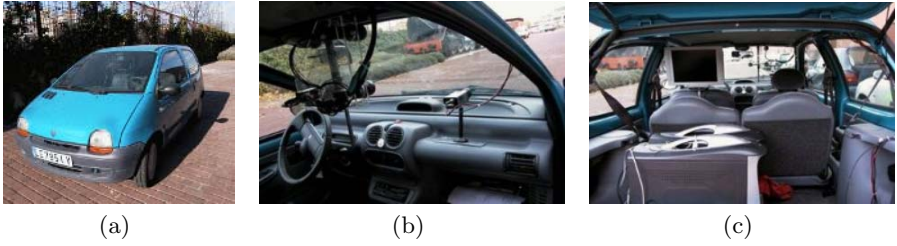


Fig. 9. (a) IvvI vehicle; (a) vision system; (c) processing system

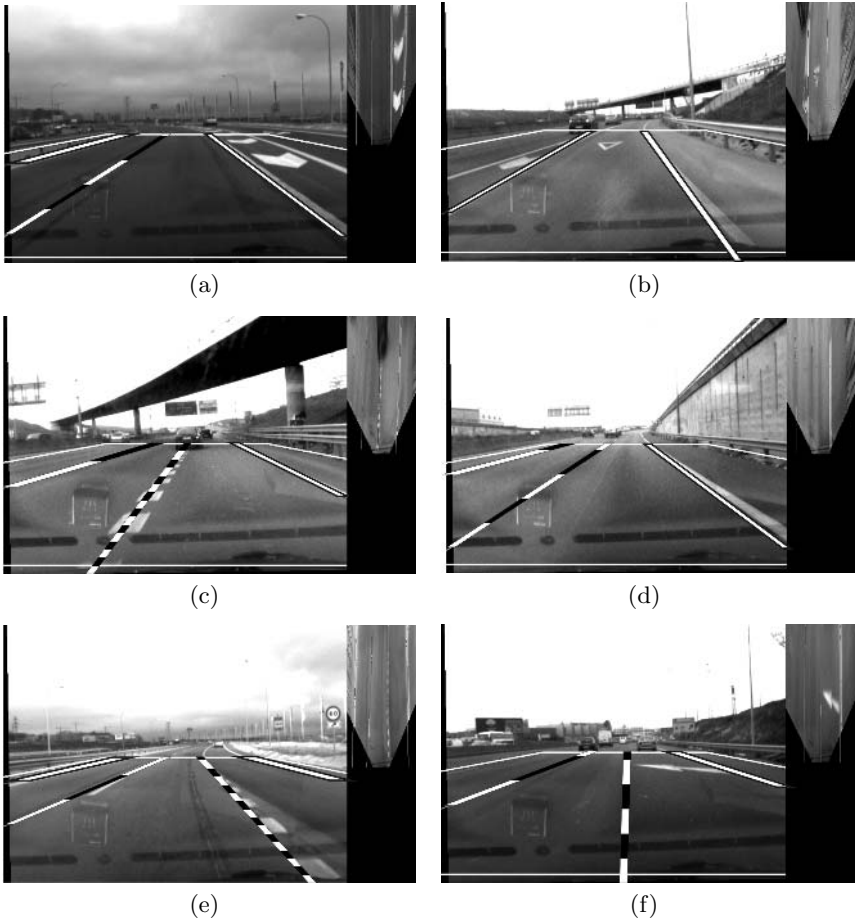


Fig. 10. Examples of detected lanes. On each example, the remapped image is at the top-right corner. The remapped region of the original image is delimited with a white line.

Figure 10 shows some examples of the execution of the algorithm. It can be seen how adjacent lanes are searched when broken or merge lane boundaries are detected.

The whole algorithm takes about 100 milliseconds in a Pentium IV, including rectification of the image – the image needs to be rectified because it comes from the left camera of a stereo-vision system –, perspective transformation (twice, due to the correction of the tilt angle), Hough Transform and road lines classification. Thus, it runs at about 10 fps. Higher rates can be achieved if the correction of the tilt angle is used to remap the next frame, instead of the current one.

4 Conclusions and Perspectives

In this paper, the Road Detection and Interpretation module of the Advanced Driver Assistance System for the IvvI project, has been presented. It is able to track the ego-lane and automatically identify lane boundary types and detect adjacent lanes if present. It can process a video sequence at nearly real time.

Detection and tracking of the road lanes is robustly performed. Also, the road line classification works reasonably good. However, this parameter of the model should be also tracked in the future, in order to filter some spurious misclassifications.

Likewise, the performance can be enhanced if interaction with other modules of the IvvI is implemented, especially with the vehicle detection one [8]. Lane position helps vehicle detection by giving an idea of the regions of the image susceptible of containing a vehicle, and the estimated size of the vehicle depending on the image position, which is related to the distance to the camera. It also helps to know if a vehicle is likely to be oncoming or out-coming depending on the lane where it is and the road type. Finally, the vehicle detection module can help the lane detection module to avoid analyzing the areas of the image occupied by other vehicles.

References

1. R. Aufrère, R. Chapuis, and F. Chausse. A model-driven approach for real-time road recognition. *Machine Vision and Applications*, 13(2):95–107, November 2001.
2. A. Broggi. Robust real-time lane and road detection in critical shadow conditions. In *IEEE International Symposium on Computer Vision*, pages 353–358, Coral Gables, Florida, November 19-21 1995. IEEE Computer Society.
3. N. W. Campbell and B. T. Thomas. Navigation of an autonomous road vehicle using lane boundary markings. In D. Charnley, editor, *1st IFAC Int. Conference on Intelligent Autonomous Vehicles*, pages 169–174. Pergamon Press, 1993.
4. R. Chapuis, R. Aufrere, and F. Chausse. Accurate road following and reconstruction by computer vision. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 3 of 4, December 2002.
5. Juan M. Collado, Cristina Hilario, Arturo de la Escalera, and Jose M. Armingol. Self-calibration of an on-board stereo-vision system for driver assistance systems. In *IEEE Intelligent Vehicle Symposium*, Tokyo, Japan, June, 13-15 2006.

6. D. DeMenthon. A zero-bank algorithm for inverse perspective of a road from a single image. In *IEEE International Conference on Robotics and Automation*, pages 1444–1449, Raleigh, NC, April 1987.
7. E. D. Dickmans and B. D. Mysliwetz. Recursive 3-d road and relative ego-state recognition. In *IEEE Transactions on Pattern Analysis and Machine Intelligence*, volume 14 of 2, pages 199–213, February 1992.
8. C. Hilario, J. M. Collado, J. M. Armingol, and A. de la Escalera. Pyramidal image analysis for vehicle detection. In *IEEE Intelligent Vehicle Symposium*, pages 87–92, Las Vegas, Nevada, U.S.A., June 6-8 2005.
9. Paul V.C. Hough. Machine analysis of bubble chamber pictures. In CERN, editor, *International Conference on High Energy Accelerators and Instrumentation*, pages 554–556, 1959.
10. M. Isard and A. Blake. Condensation – conditional density propagation for visual tracking. *International Journal of Computer Vision*, 29(1):5–28, 1998. Kluwer Academic Publishers.
11. K. Kanatani and K. Watanabe. Reconstruction of 3-d road geometry from images for autonomous land vehicles. In *IEEE Transactions on Robotics and Automation*, volume 6 of 1, February 1990.
12. K. Kluge and S. Lakshmanan. A deformable-template approach to lane detection. In *Intelligent Vehicles '95 Symposium., Proceedings of the*, pages 54–59, 25-26 September 1995.
13. R. Risack, P. Klausmann, W. Küger, and W. Enkelmann. Robust lane recognition embedded in a real-time driver assistance system. In *IEEE International Conference on Intelligent Vehicles*, pages 35–40, 1998.
14. B. Southall and C.J. Taylor. Stochastic road shape estimation. In *8th IEEE International Conference on Computer Vision (ICCV)*, volume 1, pages 205–212, 7-14 July 2001.
15. Y. Wang, D. Shen, and E.K. Teoh. Lane detection using spline model. *Pattern Recognition Letters*, 21(8):677–689, July 2000. *Pattern Recognition Letters*, vol.21, no.8, July 2000. p. 677-689.
16. Y. Wang, E. K. Teoh, and D. Shen. Lane detection and tracking using b-snake. *Image and Vision computing*, 22:269–280, July 2004.
17. A. L. Yuille and J. M. Coughlan. Fundamental limits of bayesian inference: order parameters and phase transitions for road tracking. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 22(2):160–173, February 2000.

An Approach to the Recognition of Informational Traffic Signs Based on 2-D Homography and SVMs

A. Vázquez-Reina, R.J. López-Sastre, P. Siegmann, S. Lafuente-Arroyo,
and H. Gómez-Moreno

Department of Signal Theory and Communications,
Universidad de Alcalá,

Escuela Politécnica Superior. Campus Universitario. 28805 Alcalá de Henares,
Spain

ameliovazquez@gmail.com

roberto javier.lopez@alu.uah.es

{philip.siegmann, sergio.lafuente, hilario.gomez}@uah.es

Abstract. A fast method for the recognition and classification of informational traffic signs is presented in this paper. The aim is to provide an efficient framework which could be easily used in inventory and guidance systems. The process consists of several steps which include image segmentation, sign detection and reorientation, and finally traffic sign recognition. In a first stage, a static HSI colour segmentation is performed so that possible traffic signs can be easily isolated from the rest of the scene; secondly, shape classification is carried out so as to detect square blobs from the segmented image; next, each object is reoriented through the use of a homography transformation matrix and its potential axial deformation is corrected. Finally a recursive adaptive segmentation and a SVM-based recognition framework allow us to extract each possible pictogram, icon or symbol and classify the type of the traffic sign via a voting-scheme.

1 Introduction

In this paper we handle the task of automatically detecting, recognizing, and classifying informational traffic signs. Several works have recently focused on traffic sign detection and recognition [1-9]. Some of which keep stages for sign detection and classification separated, such as [5] and [6], while some others try to address the whole process in a unique framework like [7]. Nevertheless, most of these works have only dealt with regulatory and warning traffic signs, and only a few have proposed a system to cope with guide and informational traffic signs, such as [2] and [3].

There are many challenges we must surpass in order to achieve successful results. We need to deal with some of the most common problems which usually arise in this kind of tasks, such as rotations, occlusions, variable lighting conditions of the scene, or sign deterioration. Some of these issues have been analyzed in [10]. In addition, we need to consider a great amount of different combinations of pictograms, symbols, or characters which are generally present on a typical informational traffic sign. For this

reason, we want to bring to the reader's attention that since traffic signs can usually contain variable-size text strings, and they might be present together with a very different kind of icons or pictograms (Fig. 1-a), it would be important to be able to dynamically organize hierarchically these objects in some way so we could easily perform a traffic sign classification based on this data.

Our framework is capable to overcome all these difficulties in several steps. Firstly, we detect and reorient every possible rectangular traffic sign which might be present on the scene. Subsequently, we carry out an adaptive segmentation to discriminate each character, and symbol of candidate signs from their background. Blobs are then recognized by means of a SVM framework. Due to the nature of informational traffic signs, those which resulted to be rectangular are adaptively segmented again recursively. Pictograms are then arranged vertically and horizontally. Finally the traffic sign is classified via a voting-scheme.

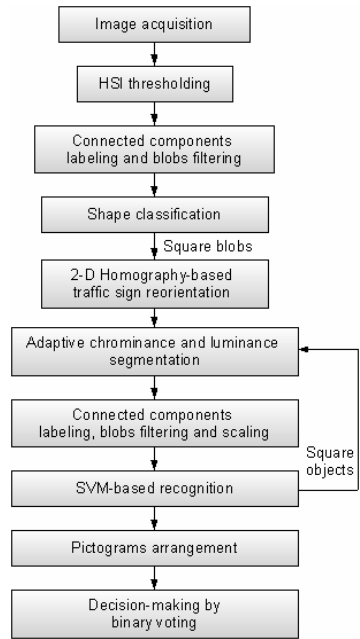
2 System Overview

Common Spanish informational traffic signs are rectangular and have a blue or white background. Foreground sign objects are designed to be clearly distinguishable from the surrounding with the help, among other things, of a high contrast to the background. Pictograms colors change generally only when they are encircled by square frames.

All these facts led us to think of dividing the process into several steps which are presented next. Initially, the original image is segmented by means of thresholding in HSI color space. This allows us to separate blue and white blobs from the context. Shape classification is then responsible for the selection of those which seem to be rectangular. Once candidate traffic signs have been extracted, we reorient them using a homography transformation matrix [11]. In the second stage we analyze the luminance and chrominance of the traffic sign in order to cope with random lighting conditions such as broad daylights, or shaded areas. Thus, we compute the colour and luminance thresholds needed to separate foreground objects from the background by way of an adaptive segmentation. This is one of the most important steps since an appropriate statistical characterization in a proper colour space may determine the success of a correct identification and recognition of every pictogram, and therefore, the right classification of the traffic sign. Afterwards, we perform connected components labeling and filter blobs in accordance with their geometrical properties such as their size or their aspect ratio. A SVM-based recognition framework classifies each blob taking as input a binary n -dimensional vector from each adaptive-segmented candidate pictogram. Blobs which are classified as square are adaptively-segmented and its pictogram classified again recursively. Objects which are successfully identified as real pictograms are then arranged vertically and horizontally by means of simple clustering, and then sorted through an adapted version of the QuickSort algorithm. A majority voting method is finally employed to get the classification from blobs position and their recognition. The complete process is outlined in Fig. 1-b.



(a)



(b)

Fig. 1. (a) Some traffic signs with several kinds of pictograms (b) System portrayal

3 Detection and Reorientation of Informational Traffic Signs

The main goal in this stage is to detect candidate traffic signs in the original scene and to reorient them. As it was mentioned above, Spanish informational traffic signs background is usually blue or white, and therefore, the first block of the detection system consists of a blue and white segmentation stage by thresholding over a given color space. We refused direct thresholding over RGB color space because, despite it might be faster under certain circumstances, it turns out to be very sensitive to lighting changes. A combination of a fixed HSI segmentation and an achromatic decomposition was consequently chosen due to its benefits as it is explained in [1].

After segmentation stage, foreground pixels are grouped together as connected components. We then classify each blob's shape employing the method described in [12] where a comparison is made between the absolute value of the FFT applied to the signature of blobs and reference shapes. Fig. 2 shows how the signature for a reference rectangular shape and for a traffic sign sample look like. 64 samples were chosen starting at 0 radians and ending at $2*\pi$ radians, and the signature was always normalized to its energy. Blobs with rectangular shape are then successfully identified and reoriented.

In the following we explain why a traffic sign reorientation is considered and why we decided to fulfill it here. First of all, a reorientation would help to make the system rotation-invariant since this would allow us to deal with only one tilt and direction regardless how they actually appear in the original scene. As long as a traffic sign may contain a lot of different kinds of icons, symbols and characters, it might result also very efficient to reorient all of them together as they theoretically should share the same tilt and distortion. Furthermore, it should be noticed that it is also much easier to gather information about objects disposition from the traffic sign vertexes rather than from each of them individually. The reorientation process is done via the Direct Linear Transformation (DLT) algorithm described in [11]. We compute H , a homography transformation matrix which univocally sets the linear relation between all the points on the reoriented traffic sign P' and on the original one P . If we consider homogeneous coordinates, given the group of points $\bar{a}_i = (x_i, y_i, z_i) \in P$ and its corresponding $\bar{a}'_i = (x'_i, y'_i, z'_i) \in P'$ we can set the following relation:

$$\bar{a}_i = H \cdot \bar{a}'_i \tag{1}$$

In a general transformation case we would have nine degrees of freedom which stand for a complete projective transformation (Fig.3-a):

$$H = \begin{bmatrix} h_{11} & h_{12} & h_{13} \\ h_{21} & h_{22} & h_{23} \\ h_{31} & h_{32} & h_{33} \end{bmatrix} \tag{2}$$

But if we consider however, a similarity transformation, results will remain practically the same as far as we suppose that traffic signs are spotted from a distance long enough when compared to their size [11]. As a result of this approximation, points coordinates from P and P' would state as $\bar{a}_i = (x_i, y_i, 1)$ and $\bar{a}'_i = (x'_i, y'_i, 1)$ respectively and we could significantly simplify our problem by reducing to four the number of variables to compute, as now, H , the homography transformation matrix, corresponds to:

$$H = \begin{bmatrix} s \cos(\theta) & -s \sin(\theta) & t_x \\ s \sin(\theta) & s \cos(\theta) & t_y \\ 0 & 0 & 1 \end{bmatrix} \tag{3}$$

Where θ denotes the rotation angle, and s , t_x and t_y represent the traffic sign scale and its translation in the X and Y axes respectively. Despite all and each of these variables define the transformation between both traffic signs P and P' in the similarity transformation case, we can not easily compute them directly from segmented blobs.

For that reason, we opted for computing H by considering a set of points correspondences which allow us to determine the four variables as in:

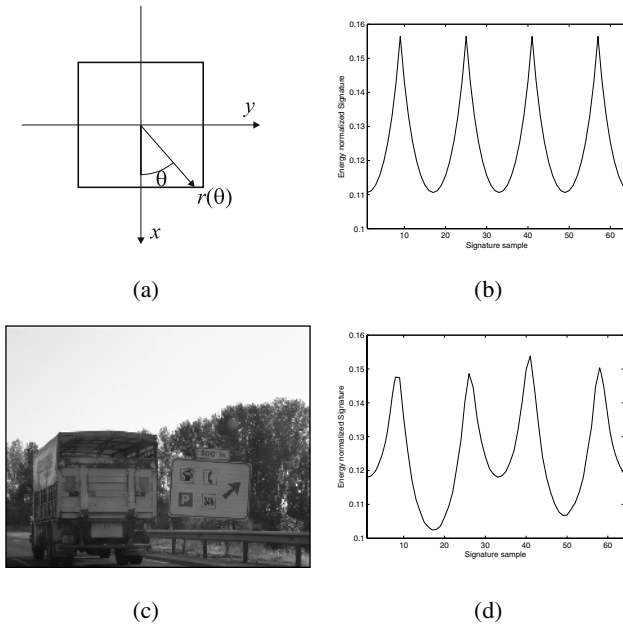


Fig. 2. Shape and traffic sign vertexes detection. (a) Ideal square blob. (b) Energy normalized signature of an ideal square blob. (c) Sample Image. (d) Energy normalized signature of Fig2-c image.

$$H = \begin{bmatrix} R & \bar{t} \\ \bar{0}^T & 1 \end{bmatrix} = \begin{bmatrix} h'_A & -h'_B & t_x \\ h_B & h'_A & t_y \\ 0 & 0 & 1 \end{bmatrix} \tag{4}$$

Theoretically we should be able to determine the four degrees of freedom of the homography matrix with two correspondences between two points each, but since this would require points coordinates to be unmistakably measured, we can rather use a greater number of correspondences so as to form an over-determined system which can be easily solved through the use of standard techniques for linear equations solving.

By reason of the former, we can use the four vertexes of the detected traffic sign already computed when calculating the blob signature to set correspondences between these four vertexes named P_1, P_2, P_3 and P_4 and those of the reoriented traffic sign we are about to get (Fig. 3-b). Finally, once H is given, we can compute each pixel of the reoriented traffic sign as:

$$\bar{x}_i = H \cdot \bar{x}'_i = R \cdot \begin{bmatrix} x'_i \\ y'_i \end{bmatrix} + \bar{t} \tag{5}$$

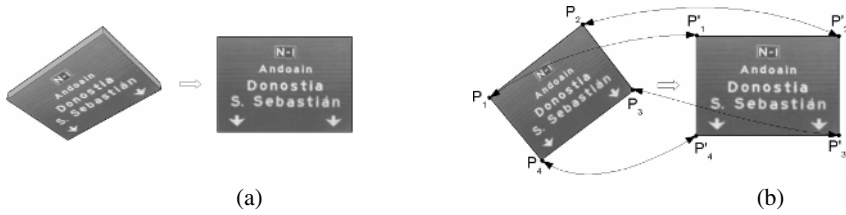


Fig. 3. (a) Projective transformation. (b) Similarity transformation and its four traffic sign vertexes correspondence.

3 Adaptive Segmentation

In order to accomplish a correct classification of an informational traffic sign, we need to know which icons and signs are actually displayed on it. These pictograms are in fact what make traffic signs different from one another, and we would like to remark that some of them may have very complex shapes. Thus, a proper segmentation of the traffic sign under test would be very convenient so that even small object details can be considered for pictograms identification.

Variable intensity conditions, the presence of noisy artifacts and possible shaded portions on a traffic sign, make very difficult to segment traffic signs in detail with fixed thresholds in a given colour space under all possible conditions. It follows that an adaptive method which might be able to extract and discriminate dynamically and accurately every object on the traffic sign would be very useful for getting fine results.

In this stage we analyze the luminance and chrominance distribution of the traffic sign in the CIE L^*a^*b color space (CIELAB from now on). CIELAB is based on the CIE 1931 XYZ colour space and consists of a luminance component and two chrominance components. It has been created to serve as a device independent model and it is considered one of the most complete colour model used to describe all the gamut of colours visible to the human eye [13]. Accuracy and efficiency discussions in the transformation from RGB to CIELAB can be found in [14].

We suppose that the amount of pixels which belong to an object background is always greater than the amount of those which pertain to an object. This fact can be noticed when computing the histogram of traffic sign chrominance or luminance components. Fig. 5 shows an example where it can be observed that there is always a wide range of values spread around a maximum peak which actually represents the most common background pixel value. Thereby, we can establish a frontier in both chrominance and luminance planes and consequently distinguish background from the foreground.

3.1 Luminance Segmentation

In the case of luminance there is only one component to work with. Supposing there is a high contrast difference between foreground and background, the function which may discriminate them is:

$$f(L) = \begin{cases} \text{“foreground/background”} & \text{if } L < \beta \\ \text{“background/foreground”} & \text{if } \beta < L \end{cases} \quad (6)$$

Where L represents the luminance component of a pixel and β represents the threshold we need to find. Depending on where the maximum peak lies, that is, in which half of the luminance histogram most common background value is located, we can distinguish which side corresponds to the foreground and which one to the background.

3.2 Chrominance Segmentation

CIELAB provide two chrominance components, and generally, the optimal chrominance function which could be able to separate the background from the foreground can be very complex and slow to evaluate. A convenient estimation can speed up the segmentation process whereas still offering good results. Since the most common background color of Spanish informational traffic signs and their respective frames and can be blue or white, we have chosen two functions $f_1(L)$ and $f_2(L)$ for evaluation which are described next.

For white backgrounds, we have $f_1(L)$ which defines a polygonal approximation of a circle C with radius r centered in the ab chrominance plane.

$$f_1(L) = \begin{cases} \text{“background”} & \text{if } (a,b) \in C_r \\ \text{“foreground”} & \text{if } (a,b) \notin C_r \end{cases} \quad (7)$$

For blue backgrounds, we have $f_2(L)$ which defines an adequate radial portion R with a proper broadness α , centered in the ab chrominance plane.

$$f_2(L) = \begin{cases} \text{“background”} & \text{if } (a,b) \in R_\alpha \\ \text{“foreground”} & \text{if } (a,b) \notin R_\alpha \end{cases} \quad (8)$$

After various experimental tests we chose the parameters α , β and r which better results offered.

4 SVM Based Recognition

Once the traffic sign has been properly segmented, we group pixels into blobs by means of connected components labeling. Each pictogram contained on the traffic sign should then result in a binary blob which will be the input to the recognition system.

The recognition framework is based on a RBF (Radial Basis Function) Kernel Support Vector Machine (SVM). SVMs are a set of related supervised learning methods which can be applied to solve many pattern recognition and regression estimation problems. They were originally introduced by Vapnik [15], [16] and they are widely used nowadays to solve binary classification problems. In these cases, if both classes could be separated by a linear hyperplane (Linear-SVMs), we would have:

- The training sets $\{x_i, y_i\}$. Where $i=1, \dots, l$, l is the number of training vectors, $y_i \in \{-1, 1\}$ identifies each class and $x_i \in \{R^d\}$ are the input feature vectors.

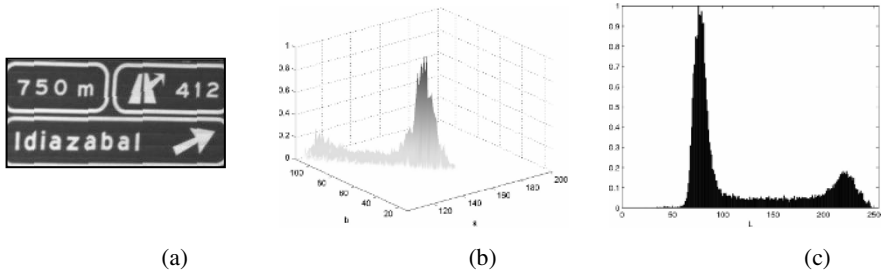


Fig. 5. (a) Informational Traffic sign which has been detected from and then reoriented. (b) Traffic sign’s chrominance distribution. (c) Traffic sign’s luminance distribution.

- The optimized hyperplane $\{w, b\}$ computed from the training sets which separates the two classes.
- The decision function given by:

$$f(x) = \text{sgn}(x \cdot w^T + b) \tag{9}$$

which determines on which side of the former hyperplane a given test vector x lies. Our case differs from the above one in two aspects. Firstly, we need to identify more than only two classes, so several one-vs-all SVMs classifiers have been actually used. Secondly, data to be classified can not usually be separated by a linear function, so we resorted to what is commonly known as the “kernel trick”. This solution consists in:

- Map the input data into a different space $\Phi(x)$ by means of the kernel function K which let us to use non-linear hyperplanes that may fit better to our problem in question.
- Build the new decision function $f(x)$ in which the scalar product of Eq .9, results in $\langle \Phi(x), \Phi(w) \rangle$, also labeled as $K(x, w)$.

$$f(x) = \text{sgn}(K(w, x) + b) \tag{10}$$

The Kernel K we chose was the RBF since it was the one which better results offered. The RBF kernel can be defined as:

$$K(x_i, x_j) = e^{-\frac{\|x_i - x_j\|^2}{2\sigma^2}} \tag{11}$$

where σ is defined as the RBF width, and x_i and x_j represent sample vectors. The SVM input vector consists of a block of 31x31 binary pixels for every candidate blob, so the interior of the bounding-box of the blob is normalized to these dimensions. σ was optimized heuristically and $\sigma = 1e-04$ was the one which better results offered. Some examples of these vectors can be seen in Fig. 6-a.

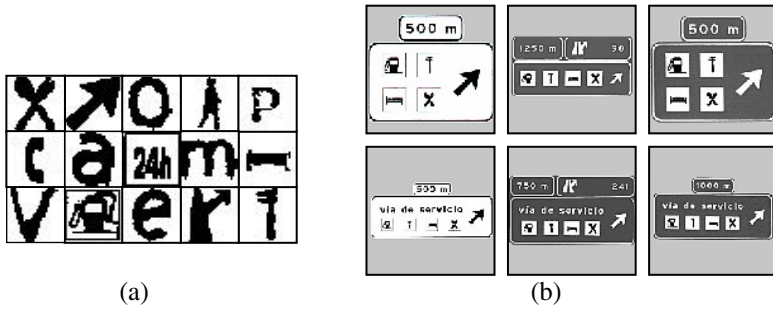


Fig. 6. (a) Sample blobs used as input vectors in the SVM recognition system. (b) Spanish Informational Traffic signs S-261, S-263, S-263a, S-264, S-266, S-266a.

5 Pictograms Arrangement

The way symbols and characters are positioned in traffic signs is not random at all, and they actually follow some fixed patterns. Pictograms relative position provide thus important information about how succesful have been the traffic sign detection and pictograms recognition. Furthermore, we can in fact gather very useful information about the type of the traffic sign to be classified from pictograms position. According to this, blobs which were succesful recognized by the SVM framework are clustered in rows and columns and sorted by means of an adapted version of Quick-Sort. Under normal circumstances no complex clustering techniques are needed since traffic signs reorientation provides enough alignment. Grid spacing is selected based on the average size of identified blobs.

7 Binary Voting

Spanish informational traffic signs differs from one another in their pictograms, characters and color schemes. Moreover, depending on the type of traffic sign, they present some specific properties which can be taken into account in a classification framework. Pictograms are usually placed following a fixed pattern which can be easily noticeable and used for identification purposes. There is also usually some redundant information which can be very useful for avoiding false alarms and making the identification more robust. Some of these examples can be seen in Fig. 6-b where Spanish traffic signs named S-261, S-263, S-263a, S-264, S-266 and S-266a [17] are showed. They all share some common properties such as an indication arrow and they differ from one another in the highway-exit pictogram and the text string “vía de servicio”. Our framework makes the most of these facts. Blobs position and identification are taken as input to a binary voting system and several conditions are setted so as to determine which traffic sign best suits to the information gathered from blobs.

8 Experimental Results

Images used for testing were compressed in JPEG. The sample set is composed of an average of 60 images for each informational traffic sign of types S-261, S-263, S-

263a, S-264, S-266 and S-266a under very different lighting conditions and environments. Tests were done in a conventional PC desktop.

Table 1 represents experimental results obtained from the above mentioned test set. A traffic sign is considered to be detected when it was properly segmented, its shape correctly classified, its blob successfully reoriented and the binary voting system recognized it as a valid informational traffic sign. False alarms occur when an image blob is wrongly considered to be a traffic sign and it is classified as one valid type of informational traffic sign. An average of 33% of false alarms was obtained from the total sample set.

Table 1. Percentage results

	S-261	S-263	S-263a	S-264	S-266	S-266a
Detection	78,00%	82,35%	86,67%	75,00%	77,78%	83,19%
Classification	72.73%	77.53%	80.31%	69.34%	75.36%	75.82%

9 Conclusions and Future Work

This paper describes a complete method to detect and recognize informational traffic signs. It is able to classify traffic signs according to their color schemes and symbols displayed on them.

The overall performance of the classifier depends mainly on how well foreground objects are extracted from the background. Chrominance and luminance analysis characterization of traffic signs and their square frames are essential, and the overall performance depends in a great extent on setting proper thresholds.

Future lines of work can include video tracking, and improvements in traffic sign detection in difficult environments. Video tracking would give more reliability to the system since more frames would be given for each traffic sign, and possible misses could be compensated with hits in other frames. Improvements in detection with shape reconstruction techniques can make the system to be able to cope with big occlusions and camera distortions.

Acknowledgment

This work was supported by the project of the Ministerio de Educación y Ciencia de España number TEC TEC2004/03511/TCM.

References

1. S. Maldonado-Bascón, S. Lafuente-Arroyo, P. Gil-Jiménez, H. Gómez-Moreno, F.López Ferreras, Road-Sign Detection and Recognition based on Support Vector Machines, IEEE Transactions on Intelligent Transportation Systems, (Submitted).
2. W. Wu, X. Chen, and J. Yang, Detection of Text on Road Signs From Video, IEEE Trans. Intelligent Transportation Systems, Vol. 6, No. 4, (2005) 378-390.

3. X. Chen, J. Yang, J. Zhang, and A. Waibel, Automatic Detection and Recognition of Signs From Natural Scenes, *IEEE Trans. Image Processing*, Vol.13 no.1, (2004) 87-89.
4. E. D. Haritaoglu and I. Haritaoglu, Real time image enhancement and segmentation for sign/text detection, in *Proc. Int. Conf. Image Processing (ICIP)*, Barcelona, Spain, vol. III, 993-996.
5. P. Paclik, J. Novovicova, P. Somol, and P. Pudill, Road sign classification using the laplace kernel classifier, *Pattern Recognition Letters*, vol. 21, (2000) 1165-1173.
6. J. Miura, T. Kanda, and Y. Shirai, An active vision system for real-time traffic sign recognition *Proc. 2000 Int Vehicles Symposium*, (2000) 52-57.
7. M. V. Shirvaikar, Automatic detection and interpretation of road signs, *Proc. of the Thirty-Sixth Southeastern Symposium on System Theory*, (2004) 413 - 416.
8. Zin, T.T.; Hama, H. Robust road sign recognition using standard deviation; 7th International IEEE Conference on Intelligent Transportation Systems, (2004) 429 - 434.
9. A. de la Escalera; J.M. Armingol, J.M. Pastor, F.J. Rodriguez; *Intelligent Transportation Systems*, *IEEE Transactions on Volume 5, Issue 2*, (2004) 57 - 68
10. S. Lafuente-Arroyo, P. Gil-Jiménez, R. Maldonado-Bascón, Traffic sign shape classification evaluation evaluation I: SVM using distance to borders, *Proc. IEEE Intelligent Vehicles Symposium*, Las Vegas, USA, (2005).
11. Hartley, R.I. and Zisserman, A. *Multiple View Geometry in Computer Vision*, Cambridge University Press, (2004).
12. P. Gil-Jiménez, S. Lafuente-Arroyo, H. Gomez-Moreno, F.López Ferreras and S. Maldonado-Bascón, Traffic sign shape classification evaluation II: FFT applied to the signature of blobs, *Proc IEEE Int Vehicles Symposium*, Las Vegas, USA, (2005).
13. G.M. Johnson and M.D. Fairchild, A top down description of S-CIELAB and CIEDE2000, *Color Research and Application*, 28, (2003) 425-435.
14. Connolly, C.; Fleiss, T. *Image Processing*, *IEEE Transactions on Image Processing*, Volume 6, Issue 7, (1997) 1046-1048.
15. V. Vapnik, *The nature of Statistical Learning Theory*. Springer-Verlog. New York, (1995).
16. V. Vapnik, *Statistical Learning Theory*, John Wiley and Sons, Inc., New York, (1998).
17. Boletín Oficial del Estado Español, Real Decreto 1428/2003, núm 306.

On Using a Dissimilarity Representation Method to Solve the Small Sample Size Problem for Face Recognition*

Sang-Woon Kim

Senior Member IEEE, Dept. of Computer Science and Engineering,
Myongji University, Yongin, 449-728 Korea
kimswo@mju.ac.kr

Abstract. For high-dimensional classification tasks such as face recognition, the number of samples is smaller than the dimensionality of the samples. In such cases, a problem encountered in Linear Discriminant Analysis-based (LDA) methods for dimension reduction is what is known as the Small Sample Size (SSS) problem. Recently, a number of approaches that attempt to solve the SSS problem have been proposed in the literature. In this paper, a different way of solving the SSS problem compared to these is proposed. It is one that employs a dissimilarity representation method where an object is represented based on the dissimilarity measures among representatives extracted from training samples instead of from the feature vector itself. Thus, by appropriately selecting representatives and by defining the dissimilarity measure, it is possible to reduce the dimensionality and achieve a better classification performance in terms of both speed and accuracy. Apart from utilizing the dissimilarity representation, in this paper simultaneously employing a fusion technique is also proposed in order to increase the classification accuracy. The rationale for this is explained in the paper. The proposed scheme is completely different from the conventional ones in terms of the computation of the transformation matrix, as well as in controlling the number of dimensions. The present experimental results, which to the best of the authors' knowledge, are the first such reported results, demonstrate that the proposed mechanism achieves nearly identical efficiency results in terms of the classification accuracy compared with the conventional LDA-extension approaches for well-known face databases involving AT&T and Yale databases.

1 Introduction

Over the past two decades, numerous families and avenues for Face Recognition (FR) systems have been developed. This development is motivated by the broad range of potential applications for such identification and verification techniques. Recent surveys are found in the literature [1] and [2] related to FR. As facial

* This work was generously supported by the Korea Research Foundation Grant funded by the Korea Government (MOEHRD-KRF-2005-042-D00265).

images are very high-dimensional, it is necessary for FR systems to reduce these dimensions. Linear Discriminant Analysis (LDA) is one of the most popular linear projection techniques for dimension reduction [3]. The major limitation when applying LDA is that it may encounter what is known as the Small Sample Size (SSS) problem. This problem arises whenever the number of samples is smaller than the dimensionality of the samples. Under these circumstances, the sample scatter matrix can become singular, and the execution of LDA may encounter computation difficulties.

In order to address the SSS issue, numerous methods have been proposed in the literature [4] - [11]. One popular approach that addresses the SSS problem is to introduce a Principal Component Analysis (PCA) step to remove the null space of the between- and within-class scatter matrices before invoking the LDA execution. However, recent research reveals that the discarded null space may contain the most significant discriminatory information. Moreover, other solutions that use the null space can also have problems. Due to insufficient training samples, it is very difficult to identify the true null eigenvalues. Since the development of the PCA+LDA [3], other methods have been proposed successively, such as the pseudo-inverse LDA [4], the regularized LDA [9], the direct LDA [5], the LDA/GSVD [11] and the LDA/QR [12]. In addition to these methods, the Discriminative Common Vector (DCV) technique [13], has recently been reported to be an extremely effective approach to dimension reduction problems. The details of these LDA-extension methods are omitted here as they are not directly related to the premise of the present work.

Recently, Duin¹ et al. [14] - [19] proposed a new paradigm to pattern classification based on the idea that if “similar” objects can be grouped together to form a class, the “class” is nothing more than a set of these similar objects. This methodology is a way of defining classifiers between the classes. It is not based on the feature measurements of the individual patterns, but rather on a suitable dissimilarity measure between them. The advantage of this is clear: As it does not operate on the class-conditional distributions, the accuracy can exceed the Bayes’ error bound². Another salient advantage of such a paradigm is that it does not have to confront the problems associated with feature spaces such as the “curse of dimensionality”, and the issue of estimating large numbers of parameters. Particularly, by selecting a set of prototypes or support vectors, the problem of dimension reduction can be drastically simplified.

¹ The authors are extremely grateful to Prof. Bob Duin for his friendly and cooperative e-mails [19], where he gave us so many helpful comments and ideas to guide the direction of the paper and to improve the quality.

² The Bayes bound can not be exceeded (by definition) in a given feature space. However, if we change representation (derived from the raw data); e.g. create better features or dissimilarities, the Bayes bound will be lower. The idea of the zero-error bound is based on the idea that dissimilarities may be defined such that there is no zero distance between objects of different classes. Consequently the classes do not overlap, and so the lower error bound is zero. We are grateful to Bob Duin for providing us with insight into this.

In this paper the utilization of the dissimilarity representation as a method for solving the SSS problem encountered in high-dimensional tasks such as face recognition is proposed. *All* samples are initially represented with the dissimilarity measures among the representatives extracted from the samples instead of the feature vectors themselves. After this transformation, an object is classified with a classifier designed in the dissimilarity space. The families of strategies investigated in this endeavor are many [18]. First in order to select such a representative set from the training set, the authors of [15] discuss a number of methods, such as random selections, the k-centers method, and others. Alternatively, investigations have centered on determining the appropriate measures of dissimilarity using measures such as various L_p Norms (including the Euclidean and $L_{0.8}$), the Hausdorff and Modified Hausdorff norm, and a number of traditional PR-based measures such as those used in template matching, and correlation-based analyses.

A problem that is encountered in this paper is one that concerns solving the SSS problem when the number of available facial images per subject is insufficient. For this reason, all of the input facial images (vectors) were selected as representatives. Following this, the dissimilarity was measured with Euclidean-based metrics with the intent of simplifying the problem for this paper. However, in facial images there are many kinds of variations, such as pose, illumination, facial expression, and distance. Thus, from simply averaging the facial images of each class, it was not possible to obtain a good representation. To overcome this problem, a combining strategy³ was employed, in which three kinds of classifiers are designed. Both of these were designed in the dissimilarity space, while the third was constructed in the input feature space. The details of these classifiers are included in the present paper.

Two modest contributions are claimed in this paper by the authors:

1. This paper lists the first reported results that reduce the dimensionality and solve the SSS problem by resorting to the dissimilarity representation. Although the result presented is only for a case when the task is face recognition, the proposed approach can also apply to other high-dimensional tasks, such as information retrieval or text classification.

2. The paper contains a formal algorithm in which two dissimilarity-based classifiers and a feature-based classifier are combined with a fusion strategy in order to improve performances for high-dimensional tasks. The paper also

³ Indeed, combination systems which fuse “pieces” of information have received considerable attention because of its potential to improve the performance of individual systems. Various fusion strategies have been proposed in the literature and workshops [20] - excellent studies are found in [21], [22], and [23]. The applications of these systems are many. For example, consider a design problem involving pattern classifiers. The basic strategy used in fusion is to solve the classification problem by designing a *set* of classifiers, and then combining the individual results obtained from these classifiers in some way to achieve reduced classification error rates. Therefore, the choice of an appropriate fusion method can further improve on the performance of the individual method, and we shall suggest methods by which we can incorporate the same ideas in our combined estimation and classification schemes.

provides an experimental comparison between this dissimilarity-based scheme and the conventional LDA-extension methods for two well-known benchmark face databases.

To the best of the authors' knowledge, all of these contributions are novel to a field of high-dimensional classification such as face recognition. This paper is organized as follows: An overview is initially presented of the dissimilarity representation in Section 2. Following this, the algorithm that solves the SSS problem by incorporating the use of dissimilarity representation and a fusion strategy is presented. Experimental results for the real-life benchmark data sets are provided in Section 3, and the paper is concluded in Section 4.

2 Dissimilarity Representation

Let $T = \{\mathbf{x}_1, \dots, \mathbf{x}_n\} \in R^p$ be a set of n feature vectors in p dimensions. It is assumed that T is a labeled data set, so that T can be decomposed into, for example, c disjoint subsets $\{T_1, \dots, T_c\}$ such that $T = \bigcup_{k=1}^c T_k, T_i \cap T_j = \phi, \forall i \neq j$. The goal here is to design a dissimilarity-based classifier in the dissimilarity space constructed with this training data set and to classify an input sample \mathbf{z} into an appropriate class.

First, for a training set of class ω_i ,

$$T_i = \{\mathbf{x}_1^{(i)}, \dots, \mathbf{x}_{n_i}^{(i)}\}, \quad \sum_{i=1}^c n_i = n, \quad (1)$$

and a representative subset extracted from T_i ,

$$Y_i = \{\mathbf{y}_1^{(i)}, \dots, \mathbf{y}_{m_i}^{(i)}\}, \quad \sum_{i=1}^c m_i = m. \quad (2)$$

As a result, $Y = \{\mathbf{y}_1, \dots, \mathbf{y}_m\}$ is obtained. After this, a dissimilarity measure, d , is additionally assumed, which is computed or derived from the samples directly. To maintain generality, a notation of $d(\mathbf{x}_i, \mathbf{y}_j)$ is used when the two samples \mathbf{x}_i and \mathbf{y}_j are quantitatively compared. Here, d is required to be nonnegative and to obey the reflexivity condition, $d(\mathbf{x}_i, \mathbf{y}_j) = 0$ if $\mathbf{x}_i = \mathbf{y}_j$, but it may be non-metric [18]. The dissimilarity computed between T and Y leads to a $n \times m$ matrix, $D(T, Y)$, in which an object \mathbf{x}_i is represented as a column vector as follows:

$$(d(\mathbf{x}_i, \mathbf{y}_1), d(\mathbf{x}_i, \mathbf{y}_2), \dots, d(\mathbf{x}_i, \mathbf{y}_m))^T, \quad 1 \leq i \leq n. \quad (3)$$

Here, the dissimilarity matrix $D(T, Y)$ is defined as a *dissimilarity space* on which the p -dimensional object given in the feature space, \mathbf{x}_i , is represented as the m -dimensional vector of Eq. (3). This column vector is simply denoted as $d(\mathbf{x})$, where \mathbf{x} is a p -dimensional vector and $d(\mathbf{x})$ is m -dimensional. From this perspective, it becomes clear that the dissimilarity representation can be considered as a *mapping* by which \mathbf{x} is translated into $d(\mathbf{x})$; thus, m is selected as sufficiently small ($m \ll p$), what is being worked in is essentially a space

with much smaller dimensions. The literature reports the use of many traditional decision classifiers, including the k -NN rule and the linear/quadratic normal-density-based classifiers to the task of classifying \mathbf{z} using $d(\mathbf{z})$ in the dissimilarity space.

2.1 Dissimilarity-Based Classifiers Without the SSS Problem

In this section, a dissimilarity-based method of classifying the high-dimensional samples without encountering the SSS problem is proposed. A Dissimilarity-Based Classifier (DBC) [24] consists of the following steps:

1. Select the representative set, Y , from the training set T by resorting to one of the prototype selection methods.
2. Compute the dissimilarity matrix, $D(T, Y)$, with T and Y , in which each individual dissimilarity is computed using one of the measures. To test a sample \mathbf{z} , compute a dissimilarity column vector, $d(\mathbf{z})$, using the same measure.
3. Achieve a classification based on invoking a classifier built in the dissimilarity space and operating on the dissimilarity vector $d(\mathbf{z})$.

However, in facial images there are many kinds of variations based on such factors as pose, illumination, facial expression, and distance. Thus, by simply measuring the differences of facial images for each class, it is not possible to obtain a good representation. To overcome this limitation, a classifier fusion strategy is employed. The basic strategy used in fusion is to solve the classification problem by designing a set of classifiers, and then to combine the individual results obtained from these classifiers in some way to achieve reduced classification error rates. The tangible rationale for this fusion strategy will be presented in a later section together with the experimental results.

The proposed approach, which is referred to as a Combined Dissimilarity-Based Classifier (CDBC), is summarized in the following:

1. Select the input training data set T as a representative subset Y ⁴.
2. Compute a dissimilarity matrix, $D(T, Y)$, by using a dissimilarity measure for all $\mathbf{x} \in T$ and $\mathbf{y} \in Y$. In addition, compute a dissimilarity column vector, $d(\mathbf{z})$, for the input sample \mathbf{z} .
3. For every class, i , perform clustering of the prototype set, Y_i into a few subsets $Y_{i,j}$, $j = 1, \dots, q_i$, $\sum_{i=1}^c q_i = q$, using any one of the clustering methods. Following this step, compute the mean vectors, $\bar{\mathbf{Y}}_{i,j}$, by averaging each cluster.
4. Perform classification of the input, $d(\mathbf{z})$ or \mathbf{z} , with *three* classifiers designed as follows:
 - (a) Classify $d(\mathbf{z})$ by invoking a k -NN classifier designed with n m -dimensional vectors in the dissimilarity space, where the prototype subset consists of each column vectors of $D(T, Y)$. The classification result is labeled as $class_1$.

⁴ This is a *Wholeset* method. Undoubtedly, for “large size” applications, other selection methods such as the *Random_C*, *KCentres*, or *PRS-based methods* [24] can be applied.

- (b) Classify $d(\mathbf{z})$ by invoking a *minimum-distance* classifier designed with c m -dimensional vectors in the dissimilarity space, where the prototype subset is obtained by averaging the column vectors of each class. The result is marked as *class₂*.
 - (c) Classify \mathbf{z} by invoking a k -NN classifier designed with q p -dimensional vectors in the input feature space, where the prototype subset is computed by averaging the mean vectors of each cluster. The result is tagged as *class₃*.
5. Obtain the final result from the *class₁*, *class₂*, and *class₃* by invoking the *Majority vote* rule, where the class which receives the largest number of votes is the *Majority* decision.

In the above algorithm, using the $n \times n$ dissimilarity matrix, the feature-based vectors are translated into the *dissimilarity-based vectors*, where the dimensionality is determined with the number of samples $n (\ll p)$. It is also interesting to note that the testing sample is projected onto the dissimilarity space represented by the dissimilarity matrix. From these considerations, it can be noted that the proposed method can be used as a scheme to reduce the dimensionality without encountering the SSS problem.

On the other hand, the above algorithm consists of three k -NN classifiers, in which the first and the second classifiers were designed in the dissimilarity space, where Euclidean distances were computed, while the third was designed in the high-dimensional feature space. Especially, the prototype mean faces for the classifier were computed in the high-dimensional space. Therefore, the distances between the prototypes may also have the problems based on the high-dimensionality, such as “the curse of dimensionality”. However, this problem could be avoided by employing a classifier design methodology by which classifiers could be designed efficiently in the dissimilarity space as well as the feature space.

The time complexity of the proposed algorithm can be analyzed as follows: Step 1 requires $O(1)$ time. Step 2 requires $O(n^2) + O(n) = O(n^2)$ time to compute the dissimilarity matrix and the dissimilarity column vector. Step 3 requires $O(c(\Gamma + q)) = O(\Gamma)$ time to cluster the training set into a small number of subsets and compute the mean vectors by averaging each cluster. Here Γ is the time for clustering each class. Step 4 requires $O(n) + O(c\tau_1) + O(q\tau_2) = O(n)$ time to classify the test sample with the *three* classifiers designed in the dissimilarity space and in the feature space. Here τ_1 and τ_2 are the times for averaging the column vectors and the mean vectors, respectively. Step 5 requires $O(1)$ time for the voting operation. Thus, the total time complexity of the CDBC is $O(n^2)$. Then the space complexity of CDBC is $O(n(n + p))$ ⁵.

⁵ In [12], it was reported that the time complexities of LDA-extension methods such as PCA, PCA+LDA, LDA/GSVD, and RLDA, respectively, are $O(n^2p)$, $O(n^2p)$, $O((n + c)^2p)$, and $O(n^2p)$ and their space complexities are all the same as $O(np)$.

3 Experimental Results

3.1 Experimental Data

The proposed method has been tested and compared with conventional methods. This was done by performing experiments on two well-known benchmark face databases, namely, the “AT&T” and “Yale” databases ⁶.

The face database captioned “AT&T”, formerly the ORL database of faces, consists of ten different images of 40 distinct subjects, for a total of 400 images. Each subject is positioned upright in front of a dark homogeneous background. The size of each image is 112×92 pixels, for a total dimensionality of 10304. The face database termed as “Yale” contains 165 gray scale images of 15 individuals. The size of each image is 243×320 pixels, for a total dimensionality of 77760. However, in this experiment, each facial image of 236×178 pixels was manually extracted, and then represented by a centered vector of normalized intensity values.

3.2 Experimental Method

In this paper, all experiments were performed using a “leave-one-out” strategy. To classify an image of object, that image is removed from the training set and the dissimilarity matrix is computed with the $n - 1$ images. Following this, all of the n images in the training set and the test object were translated into a dissimilarity space using the dissimilarity matrix, and recognition was performed based on the proposed algorithm in Section 2.1. We repeated this n times for every sample and obtained a final result by averaging them.

To construct the dissimilarity matrix, all samples were selected as representatives and the dissimilarities were measured with Euclidean Distance (ED) and Regional Distance (RD) ⁷. Here, RD is defined as the average of the minimum difference between the gray value of a pixel and the gray value of each pixel in the 5×5 neighborhood of the corresponding pixel. In this case, the regional distance compensates for a displacement of up to three pixels of the images. The details of the distance are omitted here, but can be found in the literature including [25].

Conversely, the faces for some subjects vary by pose, illumination, facial expression, and whether or not they are wearing glasses. Thus, the mean face simply obtained by averaging the input images can not work as a representative. To overcome this problem, a fusion strategy is employed in which *three* classifiers are combined. The first two were designed in the dissimilarity space, while the remaining one was constructed in the feature space. Here, to obtain the

⁶ A thorough evaluation on AT&T and Yale databases is presented here. There are more challenging datasets, such as FERET and CMU-PIE.

⁷ Here, we experimented with these simple measures, such as *ED* and *RD*. However, it should be mentioned that we can have numerous solutions, depending on dissimilarity measures, such as the modified Hausdorff distances. From this perspective, the question “what is the best measure ?” is an interesting issue for further study.

prototypes of the *third* classifier, each class was initially clustered by invoking a k-means algorithm and then averaged the clusters. To simplify the classification task for the paper, the nearest neighbor classifiers and the minimum-distance classifier were constructed. However, other classifiers, including linear/quadratic classifiers and SVM-based classifiers can also be employed.

3.3 Experimental Results

The run-time characteristics of the proposed algorithm for the two benchmark data-bases, AT&T and Yale, is reported below, and is shown in Tables 1 and 2. The performances of the dissimilarity-based classifiers (DBC and CDBC) are investigated first. Following this, a comparison is made between the conventional LDA-extension methods and the proposed CDBC scheme.

First, the numbers of the clusters in Step 3 of CDBC in Section 2.1 were probed into. Table 1 shows the classification accuracy rates (%) of CDBC for the two databases. Here, the abbreviations *ED* and *RD* in the second column indicate the dissimilarity measures employed in this experiment; specifically, the Euclidian distance and the regional distance, for the abbreviations *ED* and *RD*. Additionally, the numbers initialized with 3, 5, 7, 9, are the number of clusters for each class.

Table 1. The classification accuracy rates (%) of the Combined Dissimilarity-Based Classifiers (CDBC). Here, the values of (·) are the processing CPU-times (seconds). The details of the table are discussed in the text.

Database Names	Dissimilarity Measures	Number of Clusters per Class			
		3	5	7	9
AT&T	ED	98.75 (0.79×10^2)	98.75 (0.12×10^3)	98.75 (0.17×10^3)	99.00 (0.18×10^3)
	RD	98.75 (0.42×10^4)	99.25 (0.72×10^4)	99.00 (0.11×10^5)	99.25 (0.12×10^5)
Yale	ED	83.64 (0.11×10^3)	84.24 (0.19×10^3)	83.03 (0.27×10^3)	93.03 (0.34×10^3)
	RD	83.03 (0.63×10^4)	83.64 (0.12×10^5)	83.64 (0.18×10^5)	83.64 (0.22×10^5)

From Table 1, it is clear that the classification accuracies for the benchmark databases can be improved by increasing the number of clusters. An example of this is the classification accuracy rates (%) and the processing CPU-times (seconds) of the classifiers designed for the AT&T database measured with *RD*. The classification accuracies for the 3, 5, 7, and 9 clusters are 98.75, 98.75, 98.75, and 99.00 (%), respectively. Alternatively, the processing CPU-times (seconds) of the classifiers are 0.79×10^2 , 0.12×10^3 , 0.17×10^3 , and 0.18×10^3 , again respectively. From the table, it should be also noted that it is possible to improve the performance by effectively measuring the dissimilarity. For instance, the

classification accuracy rates of the classifiers designed with *RD* for the AT&T database are 98.75, 99.25, 99.00, and 99.25 (%), respectively. This also applies for the same characteristics for the Yale database.

Secondly, to examine the rationality of employing a fusion technique in the CDBC, the simple Dissimilarity-Based Classifier (DBC) was experimented. While CDBC involves all of the five steps given in Section 2.1, DBC consists of only the steps 1, 2, and 4(a). The classification accuracy of DBC was evaluated for the AT&T and Yale databases. In this experiment, the same dissimilarity matrix was constructed for both DBC and CDBC. Also, in both methods, each class was divided into 3, 5, 7, and 9 clusters. From the experiments, for the AT&T database, the same classification accuracies of 96.50 and 95.00 (%) were obtained throughout all the numbers of clusters for the *ED* and *RD* measures, respectively. For the Yale database, the accuracies for *ED* were found to be equal at 79.39 (%) for all the numbers of clusters, while the accuracies were also all 79.39 (%) for *RD*, regardless of the number of clusters. Thus, a comparison of Table 1 and the above results shows the fact that the proposed combined classifier, CDBC, is nearly almost always superior (and sometimes, much more superior) to the simple classifier, DBC. From this consideration, it becomes clear that the rationale of the paper for employing a fusion technique works well.

Finally, the conventional LDA-extension methods were experimented with in order to make a comparison with the CDBC. Table 2 shows a comparison of the performances of the conventional method and the proposed scheme for the two benchmark databases. Here, the PCA [3], the PCA+LDA [3], the Direct-LDA [5], the R-LDA [9], the DCV [13], and the LDA/GSVD [11] were implemented, and these methods were compared with the CDBC proposed in Section 2.1. In this experiment, in order to reduce the computational complexity, each image from the two databases AT&T and Yale, was down-sampled into 56×46 and 61×80 , respectively. It is interesting to note also that, in the PCA+LDA method, PCA was used first to reduce the dimension of the original feature space, p , to an intermediate dimension, $n - c$, where n is the total number of training samples and c is the number of classes. Secondly, LDA followed to reduce the dimension to $c - 1$ again. In the PCA method, however, the dimension p was directly reduced into $c - 1$. In this comparison, the “leave-one-out” strategy was also used to experiment with the methods. For CDBC, the dissimilarity matrix was constructed with the Euclidean distance and the number of clusters of each class was set to nine.

Initially considered are the classification accuracy rates of the AT&T database for the conventional LDA-extension methods. The accuracies of the conventional methods are, 93.25, 95.50, 98.50, 98.00, 97.25, and 93.50 (%), while the accuracy of CDBC (the Euclidean distance) is 99.00 (%). From these figures, it is apparent that the classification accuracy rate of the proposed scheme is only marginally better than those of the conventional methods. A comparison of these figures shows that the performance of the dissimilarity-based scheme works well. From Table 3, it is also noted that the processing CPU-time of feature extraction can

Table 2. A comparison of the performances of the conventional LDA- extension methods and the proposed dissimilarity-based scheme for two benchmark databases. Here, the values of (·) are the processing CPU-times. The details of the table are discussed in the text.

Types	Feature Reduction Schemes	Databases	
		AT&T	Yale
Conventional LDA-extension Methods	PCA	93.25 (0.4795×10^3)	72.73 (0.1509×10^6)
	PCA+LDA	95.50 (0.4948×10^5)	74.55 (0.1511×10^6)
	Direct-LDA	98.50 (0.2090×10^4)	92.12 (0.2785×10^3)
	R-LDA	98.00 (0.1938×10^6)	<i>Not available*</i>
	DCV*	97.25 (0.5958×10^4)	70.91 (0.1846×10^4)
	LDA/GSVD	93.50 (0.1167×10^5)	98.79 (0.2678×10^5)
Proposed Scheme	CDBC	99.00 (0.4036×10^2)	80.00 (0.2548×10^2)

be reduced significantly by employing a dissimilarity representation. It should be noted that the reduction of processing time was greatly enhanced when the dimensions of the data sets was increased.

In contrast, the classification accuracy rates for the Yale database for the conventional schemes and the proposed scheme are different from those of the AT&T database⁸. The classification accuracy rate of the dissimilarity-based scheme is poor. This undesirable result seems to originate from the fact that each image of the Yale database has a non-uniform background. As such, the dissimilarity between them was not considered for measurement. From this point of view, it appears feasible that a dissimilarity-based method can be used as a scheme for solving the SSS problem for a high-dimensional classification by developing an appropriate dissimilarity measure.

In review, it is not easy to crown any one method as superior to others for solving the SSS problem. However, as a matter of comparison, it is clear that with regard to the processing CPU-times involved, the proposed dissimilarity-based method has been shown to be computationally better than the conventional schemes.

4 Conclusions

In this paper a method that seeks to address the SSS problem by employing a dissimilarity representation method is considered. Rather than use Fisher's criterion to reduce the dimensionality, a completely different approach was employed, in which an object is represented based on the dissimilarity measures among training samples instead of from the feature vector itself. Thus, by using a small number of training samples as representatives, it was possible to reduce

⁸ In Table 2, the “*Not available*” was due to the out-of-memory problem. In order to implement the R-LDA method [9], we needed three $p \times p$ matrices. To be consistent with other methods, however, we did not fix the implementation. We also failed to obtain as good result as in [13] with DCV.

the dimensionality and achieve a better classification performance in terms of both speed and accuracy.

The proposed method has been tested on two well-known face databases and compared with conventional LDA-extensions. The experimental results demonstrate that the proposed scheme is better than conventional schemes in terms of the processing CPU-times. Although an investigation was made that focused on the possibility that the dissimilarity representation could be used to solve the SSS problem, many problems remain. This classification performance could be further improved by the development of an appropriate dissimilarity measure (i.e., the Hausdorff distance) and by the designing of suitable classifiers (i.e., linear and possibly quadratic classifiers) in the dissimilarity space. The research concerning this is a future aim of the authors.

References

1. W. Zhou, R. Chellappa, A. Rosenfeld, and P. J. Phillips.: Face recognition: a literature survey. *ACM Comput. Surveys*, **35(4)** 399–458 2003.
2. J. Ruiz-del-Solar and P. Navarrete.: Eigenspace-based face recognition: a comparative study of different approaches. *IEEE Trans. Systems, Man, and Cybernetics - Part C*, **SMC-35(3)** 315–325 2005.
3. P. N. Belhumeur, J. P. Hespanha, and D. J. Kriegman.: Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. and Machine Intell.*, **PAMI-19(7)** 711–720 1997.
4. S. Raudys and R. P. W. Duin.: On expected classification error of the Fisher linear classifier with pseudoinverse covariance matrix. *Pattern Recognition Letters*, **19** 385–392 1998.
5. H. Yu and J. Yang.: A direct LDA algorithm for high-dimensional data - with application to face recognition. *Pattern Recognition*, **34** 2067–2070 2001.
6. J. Lu, K. N. Plataniotis, and A. N. Venetsanopoulos.: Face recognition using LDA-based algorithm. *IEEE Trans. Neural Netw.*, **14(1)** 195–200 2003.
7. J. Yang, D. Zhang, and J. -Y. Yang.: A generalized K-L expansion method which can deal with small sample size and high-dimensional problems. *Pattern Analysis and Applications*, **6** 47–54 2003.
8. J. H. Friedman.: Regularized discriminant analysis. *J. Am. Statistical Assoc.*, **84(405)** 165–175 1989.
9. D. Q. Dai and P. C. Yuen.: Regularized discriminant analysis and its application to face recognition. *Pattern Recognition*, **36** 845–847 2003.
10. L. -F. Chen, H. -Y. M. Liao, M. -T. Ko, J. -C. Lin, and G. -J. Yu.: A new LDA-based face recognition system which can solve the small sample size problem. *Pattern Recognition*, **33** 1713–1726 2000.
11. P. Howland, J. Wang, and H. Park.: Solving the small sample size problem in face recognition using generalized discriminant analysis. *Pattern Recognition*, **39** 277–287 2006.
12. J. Ye and Q. Li.: A two-stage linear discriminant analysis via QR-decomposition. *IEEE Trans. Pattern Anal. and Machine Intell.*, **PAMI-27(6)** 929–941 2005.
13. H. Cevikalp, M. Neamtu, M. Wilkes, and A. Barkana.: Discriminative common vectors for face recognition. *IEEE Trans. Pattern Anal. and Machine Intell.*, **PAMI-27(1)** 4–13 2005.

14. R. P. W. Duin, D. Ridder and D. M. J. Tax.: Experiments with a featureless approach to pattern recognition. *Pattern Recognition Letters*, **18** 1159–1166 1997.
15. R. P. W. Duin, E. Pekalska and D. de Ridder.: Relational discriminant analysis. *Pattern Recognition Letters*, **20** 1175–1181 1999.
16. E. Pekalska and R. P. W. Duin.: Dissimilarity representations allow for building good classifiers. *Pattern Recognition Letters*, **23** 943–956 2002.
17. E. Pekalska.: Dissimilarity representations in pattern recognition. Concepts, theory and applications. *Ph.D. thesis, Delft University of Technology, Delft, The Netherlands*, 2005.
18. E. Pekalska, R. P. W. Duin, and P. Paclik.: Prototype selection for dissimilarity-based classifiers. *Pattern Recognition*, **39** 189–208, 2006.
19. R. P. W. Duin.: Personal communication.
20. <http://www.diee.unica.it/mcs/home.html>.
21. J. Kittler, M. Hatef, R. P. W. Duin and J. Matas.: On combining classifiers. *IEEE Trans. Pattern Anal. and Machine Intell.*, **PAMI-20(3)** 226–239 1998.
22. L. I. Kuncheva, J. C. Bezdek and R. P. W. Duin.: Decision templates for multiple classifier fusion: an experimental comparison. *Pattern Recognition*, **34** 299–414 2001.
23. L. I. Kuncheva.: A theoretical study on six classifier fusion strategies. *IEEE Trans. Pattern Anal. and Machine Intell.*, **PAMI-24(2)** 281–286 2002.
24. S. -W. Kim and B. J. Oommen.: On optimizing dissimilarity-based classification using prototype reduction schemes. (*to appear*).
25. Y. Adini, Y. Moses, and S. Ullman.: Face recognition: The problem of compensating for changes in illumination direction. *IEEE Trans. Pattern Anal. and Machine Intell.*, **PAMI-19(7)** 721–732 1997.

A Comparison of Nearest Neighbor Search Algorithms for Generic Object Recognition

Ferid Bajramovic¹, Frank Mattern^{1,*}, Nicholas Butko², and Joachim Denzler¹

¹ Chair for Computer Vision, Friedrich-Schiller-University Jena
{bajramov, mattern, denzler}@informatik.uni-jena.de
<http://www4.informatik.uni-jena.de>

² Department of Cognitive Science, University of California at San Diego
nbutko@cogsci.ucsd.edu
<http://mplab.ucsd.edu>

Abstract. The nearest neighbor (NN) classifier is well suited for generic object recognition. However, it requires storing the complete training data, and classification time is linear in the amount of data. There are several approaches to improve runtime and/or memory requirements of nearest neighbor methods: Thinning methods select and store only part of the training data for the classifier. Efficient query structures reduce query times. In this paper, we present an experimental comparison and analysis of such methods using the ETH-80 database. We evaluate the following algorithms. Thinning: condensed nearest neighbor, reduced nearest neighbor, Baram's algorithm, the Baram-RNN hybrid algorithm, Gabriel and GSASH thinning. Query structures: kd-tree and approximate nearest neighbor. For the first four thinning algorithms, we also present an extension to k -NN which allows tuning the trade-off between data reduction and classifier degradation. The experiments show that most of the above methods are well suited for generic object recognition.

1 Introduction

As shown in [1], the nearest neighbor classifier works well for generic object recognition. However, a naive implementation requires storing the complete training set, and classification takes time proportional to the size of the training data times the dimension of the feature vectors. Both aspects can be improved: efficient query structures greatly reduce classification time and thinning methods reduce the amount of data which has to be stored for the classifier. In this paper, we evaluate the performance of several such methods: for classification, we use kd-trees and approximate nearest neighbor (ANN), and for thinning, condensed nearest neighbor (CNN), reduced nearest neighbor (RNN), Baram's algorithm, Baram-RNN hybrid algorithm, Gabriel and GSASH thinning. For CNN, RNN, Baram and Baram-RNN, we propose and evaluate an extension to k nearest neighbors, which allows tuning the extent of thinning and thus the trade-off between data reduction and degradation of classification rates.

* This work was financially supported by the German Science Foundation (DFG), grant no. DE 732/2-1.

The remainder of the paper is organized as follows: In section 2 we give a short repetition of the k nearest neighbor classifier. Sections 3 and 4 describe the efficient query structures and the thinning algorithms respectively. In section 5 we present our experimental results. Section 6 gives final conclusions.

2 Nearest Neighbor Classifier

The k nearest neighbors (k -NN) classifier requires a labeled training data set $\{\mathcal{X}, \mathcal{Y}\} = \{(\mathbf{x}_1, y_1), \dots, (\mathbf{x}_n, y_n)\}$ consisting of d dimensional feature vectors \mathbf{x}_i and their class labels y_i . For $k = 1$, in order to classify a new feature vector \mathbf{x} , find the closest element \mathbf{x}_i in \mathcal{X} and assign the label y_i to \mathbf{x} . The misclassification error of the 1-NN classifier converges (for $n \rightarrow \infty$) to at most twice the Bayes-optimal error [2].

For $k > 1$, find the k nearest neighbors $(\mathbf{x}_{i_1}, \dots, \mathbf{x}_{i_k})$ of \mathbf{x} in \mathcal{X} . Then perform a voting amongst the class labels $(y_{i_1}, \dots, y_{i_k})$ of these neighbors. The classic rule is to choose the class with the most votes within the set of neighbors, breaking ties arbitrarily. For $k > 1$, the asymptotic ($n \rightarrow \infty$) misclassification error of the k -NN classifier is as low as the Bayes-optimal error [2]. The voting can be modified to include a rejection rule. There are several possibilities: reject ties, reject if majority is too small, reject if not all neighbors are in the same class (unanimous voting). In general, the stricter the voting rule is, the more rejections there will be, but also the lower the misclassification rate will be.

3 Efficient Query Structures

There are several approaches to improve the running time of brute force nearest neighbor search [3,4,5]. In higher dimensions, however, these algorithms have an exponentially growing space requirement. Besides the small asymptotic improvement in time which was achieved by Yao and Yao [6] there exists no exact algorithm which can improve both time and space requirements in the worst case.

3.1 kd-Tree

The practically most relevant approach known for higher dimensions is the kd-tree introduced by Friedman, Bentley and Finkel [7]. The idea of the kd-tree is to partition the space using hyperplanes orthogonal to the coordinate axes. Each leaf node contains a bucket with a number of vectors, the other nodes in the binary kd-tree consist of a splitting dimension d and a splitting value v .

A query only has to look at one dimension of the query point at each node to decide into which subtree to descend. After the closest vector \mathbf{x} in the bucket is found, one also has to search all buckets which are closer to the query vector than \mathbf{x} . In order to keep the tree small and to avoid searching in many buckets, one can stop splitting the tree if the bucket has a reasonably small size and search in the bucket linearly. It is shown in section 5 that this can reduce query times.

If the data is organized in a balanced binary tree, running time in the expected case is logarithmic. Unfortunately, the running time depends on the distribution of the training

data. In the worst case, the running time is linear. To improve the running time, several splitting rules were defined by [8].

The *standard kd-tree splitting rule* chooses the dimension as splitting dimension in which the data \mathcal{X} have maximum spread. The splitting threshold is the median of the coordinates of \mathcal{X} along this dimension. The depths of the tree is ensured to be $\lceil \log_2 n \rceil$. But theoretically, the bucket cells can have arbitrarily high aspect ratio.

The *midpoint splitting rule* guarantees cells with bounded aspect ratio. It cuts the cells through the mean of its longest side breaking ties by choosing the dimension with maximum spread. *Trivial splits*, where all vectors of \mathcal{X} lie on one side of the splitting plane, can occur and possibly cause the tree to have a larger depth than n .

The *sliding-midpoint splitting rule* is defined as the midpoint splitting rule, but omits trivial splits by replacing such a split with a split which contains at least one vector on each side. This is achieved by moving the splitting plane from the actual position up to the first vector of the dataset. This ensures that the maximum possible tree depth is n .

The *fair-split rule* is a compromise between the standard and midpoint splitting rules. The splitting plane is chosen from the possible coordinates in which a midpoint split can be done that does not exceed a certain aspect ratio of longest to shortest side. Among these, the coordinate with the maximum spread is chosen. The two extreme splitting planes which fulfill the aspect ratio will be compared with the median of the coordinates. If the median is on the smaller side, the cut will be done. Otherwise, a cut will be done at the median. Again, trivial splits can cause the tree depth to exceed n .

The *sliding fair-split rule* works as the fair-split rule but omits empty buckets by considering the extreme cut which just does not exceed a certain aspect ratio and which is closer to the median if the median does not fulfill the aspect ratio criterion. If this extreme cut is a trivial one, it is moved up to the position such that one vector lies on the other side. Again, this ensures that the maximum depth of the tree is n .

3.2 kd-Tree for Approximate Nearest Neighbor

Applying NN classification to generic object recognition, it is not important to really find the nearest neighbor. The classification is correct if a datapoint of the same class is found. So we consider doing generic object recognition with an approximate nearest neighbor approach developed by Arya and Mount [9]. A $(1 + \epsilon)$ approximate nearest neighbor is defined as follows:

Definition 1. A vector \mathbf{q} is called $(1 + \epsilon)$ approximate nearest neighbor of $\mathbf{x} \in \mathcal{X}$ if for all $\mathbf{y} \in \mathcal{X}$: $d(\mathbf{x}, \mathbf{q}) \leq (1 + \epsilon)d(\mathbf{y}, \mathbf{q})$.

The value ϵ is also called the error bound. If $\epsilon = 0$, the query is equivalent to the exact nearest neighbor classification. Otherwise, the minimum distance to the real nearest neighbor is at least $1/(1 + \epsilon)$ of the found distance.

To find a given query vector \mathbf{q} , the leaf cell in the tree is found by descending the tree. Only those neighboring cells which are in the range of $d(\mathbf{x}, \mathbf{q})/(1 + \epsilon)$ are searched for a closer training vector. Arya [9,10] has shown that the algorithm has polylogarithmic query time and needs nearly linear space which can be made quite independent of the vector distribution.

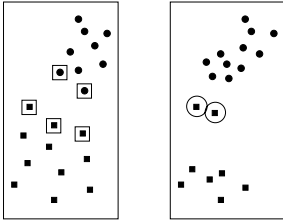


Fig. 1. Both images show the training data of two class problems. In the left image, the subset indicated by the boxes is a 2-consistent subset. The set in the right image is 2-consistent, but not 3-consistent, because the vectors indicated by the circles are 3-inconsistent.

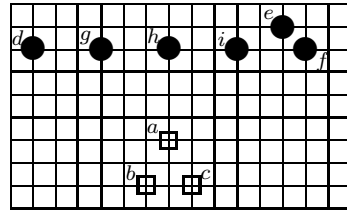


Fig. 2. The image shows a 3-consistent two class training set, which can be thinned by a 3-NN condensed nearest neighbor thinning to a 3-inconsistent set. The example uses the manhattan distance. If the vectors are visited in the order $a, b, c, d, e, f, g, h, i$, all vectors except for i will be added to the thinned set. In this set, h is 3-inconsistent.

4 Thinning

Thinning means reducing the training data set $\{\mathcal{X}, \mathcal{Y}\}$ to a smaller subset $\{\mathcal{X}', \mathcal{Y}'\}$. The classifier then only uses $\{\mathcal{X}', \mathcal{Y}'\}$. This results in reduced memory requirements and query times. There is an important property of thinned data sets $\{\mathcal{X}', \mathcal{Y}'\}$ [2]:

Definition 2. A set $\{\mathcal{X}', \mathcal{Y}'\} \subseteq \{\mathcal{X}, \mathcal{Y}\}$ is called consistent subset of $\{\mathcal{X}, \mathcal{Y}\}$ if the 1-NN classifier for $\{\mathcal{X}', \mathcal{Y}'\}$ correctly classifies all members of the original set $\{\mathcal{X}, \mathcal{Y}\}$.

This property is very desirable, as it guarantees perfect recognition of the 1-NN classifier for $\{\mathcal{X}', \mathcal{Y}'\}$ applied to the whole training set $\{\mathcal{X}, \mathcal{Y}\}$. We extend the definition with respect to the k -NN classifier:

Definition 3. A vector $x \in \mathcal{X}$ is called k -consistent with respect to $\{\mathcal{X}, \mathcal{Y}\}$ if the unanimous k -NN classifier for $\{\mathcal{X}, \mathcal{Y}\}$ classifies it correctly. Otherwise it is called k -inconsistent with respect to $\{\mathcal{X}, \mathcal{Y}\}$. A set $\{\mathcal{X}, \mathcal{Y}\}$ is called k -consistent set if it has no elements which are k -inconsistent with respect to $\{\mathcal{X}, \mathcal{Y}\}$. A subset $\{\mathcal{X}', \mathcal{Y}'\} \subseteq \{\mathcal{X}, \mathcal{Y}\}$ is called k -consistent subset of $\{\mathcal{X}, \mathcal{Y}\}$ if all members of $\{\mathcal{X}, \mathcal{Y}\}$ are k -consistent with respect to $\{\mathcal{X}', \mathcal{Y}'\}$.

Clearly, the terms consistent subset and 1-consistent subset are equivalent. As for the 1-NN case, the property k -consistent subset guarantees perfect recognition of the k -NN classifier for $\{\mathcal{X}', \mathcal{Y}'\}$ applied to the whole training set $\{\mathcal{X}, \mathcal{Y}\}$. Fig. 1 shows an example of a 2-consistent subset for a given training set. Next, we proof three theorems:

Theorem 1. A vector which is k -consistent with respect to a set is also k' -consistent with respect to the same set, for all $k' \leq k$.

Proof: The k nearest neighbors of a labeled vector (x, c) , which is k -consistent with respect to $\{\mathcal{X}, \mathcal{Y}\}$, are all in class c . Thus, also its k' nearest neighbors are in class c . Thus, (x, c) is k' -consistent with respect to $\{\mathcal{X}, \mathcal{Y}\}$. □

Input: $\{\mathcal{X}, \mathcal{Y}\}$	
Initialize R with one random element of $\{\mathcal{X}, \mathcal{Y}\}$	
FOR EACH $(x, c) \in \{\mathcal{X}, \mathcal{Y}\} \setminus R$	
IF	x is k' -inconsistent with respect to R
THEN	Set $R = R \cup (x, c)$
UNTIL R has not changed during the previous FOR EACH loop	
Result: $\{\mathcal{X}', \mathcal{Y}'\} = R$	

Fig. 3. Hart’s thinning algorithm: condensed nearest neighbor

Theorem 2. *A k -consistent subset of a set is a k -consistent set.*

Proof: Given a k -consistent subset $\{\mathcal{X}', \mathcal{Y}'\}$ of $\{\mathcal{X}, \mathcal{Y}\}$, all elements of $\{\mathcal{X}, \mathcal{Y}\}$ are k -consistent with respect to $\{\mathcal{X}', \mathcal{Y}'\}$. As $\{\mathcal{X}', \mathcal{Y}'\} \subseteq \{\mathcal{X}, \mathcal{Y}\}$, all elements of $\{\mathcal{X}', \mathcal{Y}'\}$ are k -consistent with respect to $\{\mathcal{X}', \mathcal{Y}'\}$. Thus, $\{\mathcal{X}', \mathcal{Y}'\}$ is a k -consistent set. \square

Theorem 3. *A k -consistent subset of a set is a k' -consistent subset of the same set, for all $k' \leq k$.*

Proof: Let $\{\mathcal{X}', \mathcal{Y}'\}$ be a k -consistent subset of $\{\mathcal{X}, \mathcal{Y}\}$. All labeled vectors in $\{\mathcal{X}, \mathcal{Y}\}$ are by definition k -consistent with respect to $\{\mathcal{X}', \mathcal{Y}'\}$ and thus k' -consistent with respect to $\{\mathcal{X}', \mathcal{Y}'\}$ (theorem 1). Thus, $\{\mathcal{X}', \mathcal{Y}'\}$ is a k' -consistent subset of $\{\mathcal{X}, \mathcal{Y}\}$. \square

4.1 Condensed Nearest Neighbor

Hart [2,11] proposed a thinning algorithm called condensed nearest neighbor (CNN). First, one element of the training set is chosen arbitrarily. Then, a scan over all remaining elements is performed. During the scan, all elements which are 1-inconsistent with respect to the new growing set are added to the new set. Additional scans are performed until the new set does not change during a complete scan. The thinned subset is guaranteed to be a 1-consistent subset of the training set [2].

While Hart’s algorithm reduces the size of the data and thus improves memory requirements and query times, it typically also reduces the recognition rate [2]. Depending on the structure of the training data and the application, the degradation of the classifier may be unacceptable. Hence, we propose an extension to the algorithm. The only change is that we require vectors to be k' -consistent instead of only 1-consistent. The complete algorithm is given in Fig. 3. The runtime of a naive implementation is $O((d + k')n^3)$ in the worst case.

While the thinned set is not guaranteed to be a k' -consistent set, as can be seen from the counter example in Fig. 2, it is obviously guaranteed to be a 1-consistent subset of the training set. It is quite obvious from theorem 3 in combination with the growing nature of the algorithm that in general, a greater value of parameter k' will result in a greater thinned set. The second part of our proposal is to choose $k' \geq k$. This means that we use a greater (or equal) parameter k' for thinning than for the application of the k -NN classifier. On the one hand, this makes sense, because even for a k' -consistent training set, the thinned subset is not guaranteed to be k' -consistent, but with increasing k' , the chances of the thinned subset being at least k -consistent increase. On the other

Input: training data $\{\mathcal{X}, \mathcal{Y}\}$ and thinned data $\{\mathcal{X}', \mathcal{Y}'\}$	
Set $R = \{\mathcal{X}', \mathcal{Y}'\}$	
FOR EACH $(x, c) \in R$	
IF	All $(x', c') \in \{\mathcal{X}, \mathcal{Y}\}$ are k' -consistent with respect to $R \setminus \{(x, c)\}$
THEN	Set $R = R \setminus \{(x, c)\}$
Result: $\{\mathcal{X}'', \mathcal{Y}''\} = R$	

Fig. 4. Postprocessing algorithm for reduced nearest neighbor

hand, while it is desirable to have a k -consistent set (or even better, a k -consistent subset of the training set), what is more important is the classification rate of the k -NN classifier for the thinned set on a separate test set. Thus, it makes perfect sense to choose $k' > 1$ for a 1-NN classifier, even though the thinned set is already guaranteed to be a 1-consistent subset of the training set for $k' = 1$. To summarize, the parameter k' can be used to tune the trade-off between data reduction and classifier degradation.

4.2 Reduced Nearest Neighbor

Gates [2,12] proposed a postprocessing step for the CNN thinning algorithm. As the initial members of the thinned set are chosen arbitrarily and as additional members are added, it may be possible to remove some vectors and still retain a 1-NN consistent subset of the training set. The postprocessing algorithm simply checks for each vector of the thinned set if the thinned set without that vector is still a 1-NN consistent subset of the training set. If it is, the vector is removed. Of course this algorithm can also be extended to a k' -NN version as described in the previous subsection. The postprocessing algorithm is given in Fig. 4. The runtime of a naive implementation is $O((d + k')n^3)$ in the worst case. The complete reduced nearest neighbor (RNN) thinning algorithm performs CNN thinning followed by the postprocessing algorithm. As the CNN part may produce a k' -inconsistent set and the postprocessing will not remove any k' -inconsistent vectors, the RNN thinning algorithm can also produce a k' -inconsistent set.

4.3 Baram's Method

Baram [2,13] proposed a thinning algorithm that thins each class individually. For each class, a new set for the thinned class is initialized with an arbitrary member of that class. Then, each vector of that class, which is 1-inconsistent with respect to a modified training set in which the current class is replaced by the growing thinned version of that class, is added. Naturally, also this algorithm can be extended to a k' -NN version. Fig. 5 shows the complete algorithm. The k' -NN version of Baram's algorithm can also produce a k' -inconsistent set, as the same counter example as for CNN applies (Fig. 2). The runtime of a naive implementation is $O((d + k')n^2)$ in the worst case. In an unpublished paper, Olorunleke [14] proposed combining Baram's algorithms with the postprocessing step of RNN and calls it Baram-RNN hybrid algorithm.

Input: $\{\mathcal{X}, \mathcal{Y}\}$	
FOR EACH class $c \in \mathcal{Y}$	
Remove class c : $\{\mathcal{X}^*, \mathcal{Y}^*\} = \{\mathcal{X}, \mathcal{Y}\} \setminus \{\mathcal{X}_c, \mathcal{Y}_c\}$	
Set $R_c = \emptyset$	
FOR EACH vector $\mathbf{x} \in \mathcal{X}_c$	
IF	(\mathbf{x}, c) is k' -inconsistent with respect to $\{\mathcal{X}^*, \mathcal{Y}^*\} \cup R_c$
THEN	Set $R_c = R_c \cup (\mathbf{x}, c)$
Result: $\{\mathcal{X}', \mathcal{Y}'\} = \bigcup_{c \in \mathcal{Y}} R_c$	

Fig. 5. Baram’s thinning algorithm ($\{\mathcal{X}_c, \mathcal{Y}_c\} \subseteq \{\mathcal{X}, \mathcal{Y}\}$ contains the members of class c)

4.4 Proximity Graph Based Thinning

The thinning algorithms in the previous sections all exhibit the property that different thinned-sets will result from considering the datapoints in a different order. As this is undesirable, we also consider order-independent, graph-based thinning algorithms.

The starting place for these order-independent algorithms is the *Delaunay graph* [15], which is constructed by connecting nodes in adjacent Voronoi cells. A Voronoi cell is the region of space around a point that is closer to that point than to any other point. If we remove a point from our set, all points falling in its Voronoi cell will now fall in a cell belonging to one of its neighbors in the Delaunay graph. This suggests a thinning algorithm: by removing all points that are surrounded by Delaunay neighbors of the same class, we are left with a thinned set that has exactly the same classification properties as the original set in a 1-NN classification scheme.

Despite its desirable properties, Delaunay Graph thinning has two critical drawbacks: the algorithm is exponential in the dimensionality of the data, and empirically removes very few points for real datasets [15]. It seems that tolerating some shift in the decision boundary can (greatly) increase the number of points removed in thinning.

Two points $\mathbf{x}_1, \mathbf{x}_2 \in \mathcal{X}$ are Gabriel neighbors if there is no third point $\mathbf{x}_3 \in \mathcal{X}$ inside the hypersphere centered on the midpoint of \mathbf{x}_1 and \mathbf{x}_2 , and with diameter equal to the distance between them. Mathematically, we say that \mathbf{x}_1 and \mathbf{x}_2 are Gabriel neighbors iff $\forall \mathbf{x}_3 \in \mathcal{X}, d(\mathbf{x}_1, \mathbf{x}_2)^2 \leq d(\mathbf{x}_1, \mathbf{x}_3)^2 + d(\mathbf{x}_2, \mathbf{x}_3)^2$. A *Gabriel graph* is an undirected graph built by connecting each node to all of its Gabriel neighbors. As with Delaunay graphs, we will consider a thinning algorithm in which all points that are only neighbors with points of the same class are removed from the dataset. Since the Gabriel graph is a subset of the Delaunay graph, Gabriel thinning will remove all of the points that Delaunay thinning removes, and possibly more. This may change the decision boundary (and possibly even leads to a 1-inconsistent thinned subset), but in practice, Sánchez *et al.* found that Gabriel thinning leads to better classification (at the cost of keeping more points) than traditional CNN methods [16].

There is a quadratic cost to finding a given point’s Gabriel neighbors. To build an entire graph so that we can do filtering, we incur this cost for every point in the data set. This means that building an exact Gabriel graph is cubic in the number of data points, and so is very costly. Using Mukherjee’s *GSASH* data structure [17], the cost becomes $O(n \log_2 n)$, though with potentially large constants.

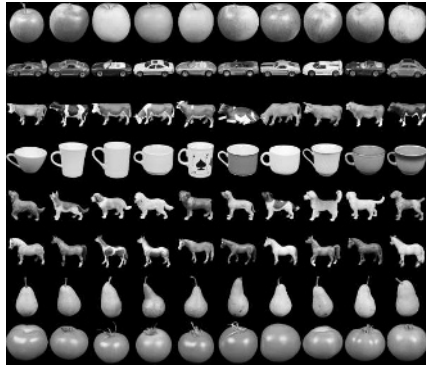


Fig. 6. Sample images of the objects of the ETH-80 database [18]

5 Results

5.1 Dataset

To evaluate the nearest neighbor classification for generic objects, we use the ETH-80 database [18], which contains 80 objects from 8 categories: apple, car, cow, cup, dog, horse, pear and tomato (see Fig. 6). Each object is represented by 41 images of views from the upper hemisphere. The experiments are performed using 128×128 pixel images, with each image cropped close to the object boundaries. The grayvalues of the image will be transformed to a feature vector by a PCA transformation with the eigenvectors of the 100 largest eigenvalues. The 100 dimensional feature vectors will be used for classification. The test is performed by cross-validation with a leave-one-object-out strategy. One of the 80 objects is used for testing and the 79 other objects are used to learn the PCA transformation and build the k -NN query structure. This “unknown” object must accordingly be classified into the correct object category.

5.2 Experiments

We examined the presented methods with respect to query time, error / rejection / recognition rate and used data size. In many applications, NN is not applied because it is too slow. This can be improved with the kd-tree. Arya [10] has shown the dependency on the splitting rule.

One other important parameter of the kd-tree is the bucket size. If it is too small, the tree becomes very large and the search for the bucket which has to be taken into consideration takes long. If the bucket size is too large, it takes too much time to search the bucket linearly. To get a fast kd-tree query, the optimal bucket size for the generic ETH-80 dataset should be medium size. In our example, bucket size 32 with the standard kd-tree splitting rule is the best choice. The splitting rule is not very important if the bucket size is chosen well. For this bucket size, query times vary by about 8% from $88.9 \mu\text{s}$ with the standard splitting rule to $96.1 \mu\text{s}$ with the midpoint splitting rule,

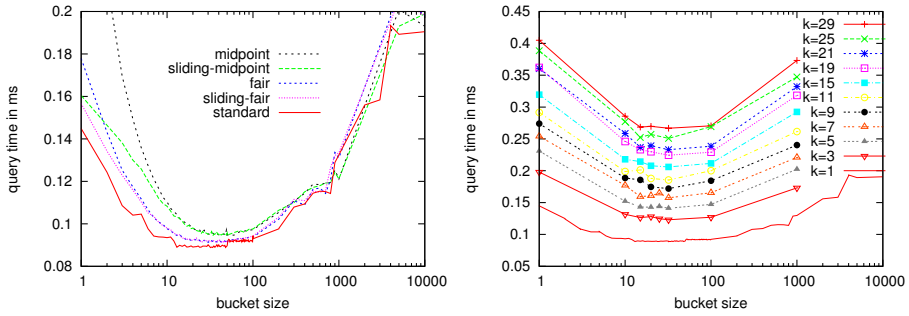


Fig. 7. Dependency of query time on a Intel Pentium 4 with 3.4GHz on a kd-tree with different bucket sizes using the generic ETH-80 dataset. Different splitting rules for $k = 1$ (left) and different parameters k for the k -NN classifier with standard splitting rule (right) are examined. Best query times are achieved with bucket size between 20 and 32.

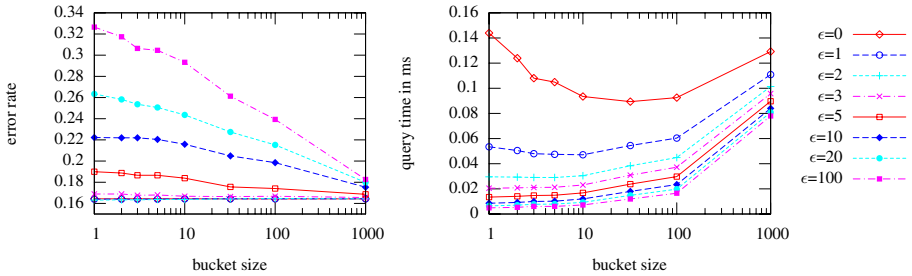


Fig. 8. Performance on different approximation error bounds ϵ . The left image shows that the error rate on smaller bucket sizes increases, but as the right image shows the query time can become amazingly fast. For $\epsilon = 100$ the query time is $5\mu s$ but causing an error rate of 32.7% with a 1-NN classifier on the generic ETH-80 dataset.

whereas with bucket size one, query times vary by a factor of about two. Using a larger parameter k for the k -NN classification, the query time increases, but the best query time is still attained at the same bucket sizes. So this is independent of k (see Fig. 7).

The query time can further be decreased by using approximate nearest neighbor classification. In general, the query time decreases with larger error bounds and also with lower bucket sizes if the error bound is large enough. In our experiments, the best query time ($5\mu s$) is obtained using $\epsilon = 100$ and bucket size one, but at the cost of a strong rise of the error rate to 32.7%. Useful values of ϵ are about 1–3 (see Fig. 8). Using an error bound $\epsilon = 2$, the query time can be improved by the factor of 3 to $29.0\mu s$ without losing any recognition rate in our 80 test sets and with $\epsilon = 3$ to $20.5\mu s$ with an increased error rate of 0.46 percentage points, which is quite acceptable.

Gabriel thinning reduces the data set only to 96.8%. The fastest and least precise GSASH approximation with one parent and one child reduces the data set to 93.8% and with 6 children and 6 parents to 94.6%. So the results are similar to those using the

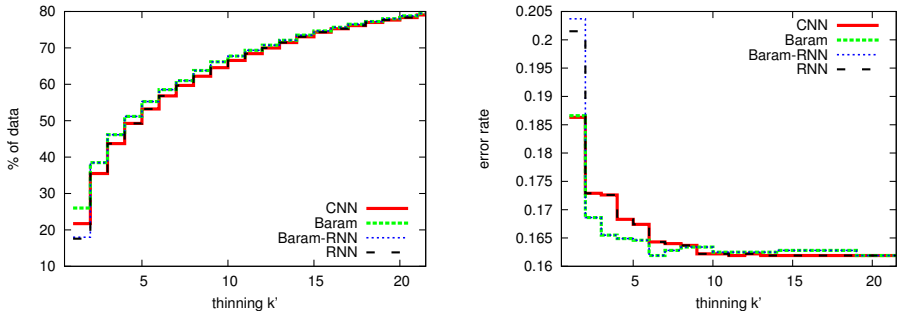


Fig. 9. In the images, the dependency of different thinning algorithms on the parameter k' is presented. The left image shows the proportion of the full dataset which remains after thinning. In the right image, the error rate using 1-NN on the reduced dataset is shown.

original data set, but can also improve the query time to error rate ratio (see Fig. 12). A major disadvantage of the approach is the time requirement for thinning. Gabriel thinning needs about 27 minutes for 3239 vectors and the fastest approximation about 11 minutes, whereas e.g. Baram or CNN need only about 2.5 seconds. The reason for the small reduction is an indication of the bad distribution of the data in the 100 dimensional space. A reduction of the dataset can still be done with CNN or Baram but at the cost of recognition rate. Our extension of the NN thinning algorithms can adjust the reduction of the dataset. This effect can be observed in Fig. 9 (left).

As expected, the greater the value of k' is, the more data is retained after thinning. Accordingly, as can be seen in Fig. 9 (right), the error rate decreases with increasing k' . Furthermore, Fig. 10 shows the influence of k' on the recognition and rejection rates for Baram and several values of k (unanimous voting). Independent of k and in accordance with Fig. 9, the recognition properties of the classifier improve with growing k' . This shows that the trade-off between data size and error rate can be tuned.

Considering Fig. 9 again, a comparison between the four thinning algorithms shows that for $k' > 1$, a) there is no difference between CNN and RNN as well as between Baram and Baram-RNN and b) Baram keeps a bit more data than CNN. While it is not surprising that hence the error rate of Baram is lower, it is actually lower than the sheer amount of data would suggest: Baram with $k' = 6$ keeps 58.5% of the data while CNN with $k' = 9$ retains 64.6%. The error rate is 16.2% in both cases. On this data set, for $1 < k' < 9$, Baram clearly outperforms CNN.

Varying parameter k for k -NN classification with unanimous voting lets us choose a specific error rate versus rejection rate ratio (see Fig. 11). If k becomes larger, the error rate decrease, but the rejection rate increase. Thinning with $k' < k$ does not make sense, because the k -NN rejection rate strongly increases, as Fig. 10 shows. Thinning with $k' = 1$ is, with respect to recognition rate, bad in general. The error rate versus rejection rate ratio is vitally better for k -NN trained with the full data set. Using Baram thinning with $k' = 6$ reaches the best possible ratio – even better than without thinning.

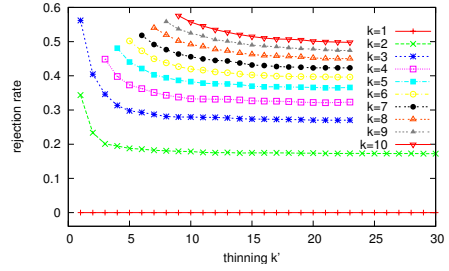
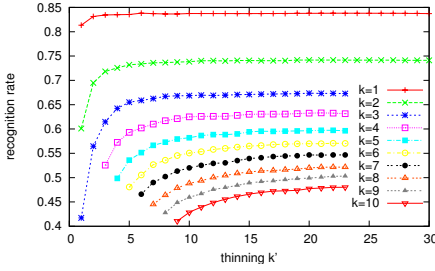


Fig. 10. Error versus rejection rate of Baram using different parameter k' for thinning and different parameter k for classification with unanimous voting

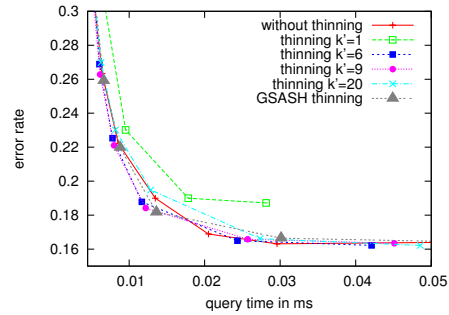
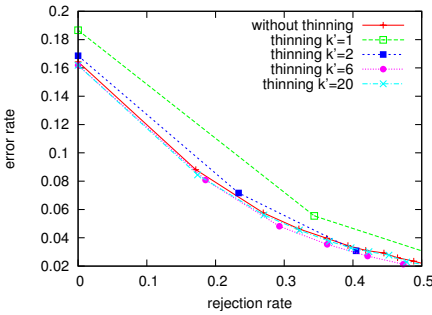


Fig. 11. Rejection versus error rate of k -NN classification (varying k) with unanimous voting on training data thinned by Baram using several parameters k'

Fig. 12. Query time versus error rate of 1-ANN classification for different parameters ϵ and training data thinned by Baram with different parameters k' and also GSASH thinning

Using higher approximation or smaller training data for nearest neighbor classification leads to a higher error rate. Which ratio between query time and error should be chosen highly depends on the application. The best methods at a given query time form an optimal ratio curve. As shown in Fig. 12, 1-ANN classification trained with data thinned by Baram with $k' = 1$ is in general worse than 1-ANN trained with the original data. A smaller error rate with respect to a given query time can be obtained using Baram thinning with e.g. $k' = 9$. Using these Baram thinned data, which are reduced to 66.2% of the original data, the ANN classification with $\epsilon = 5$ attains a query time of 12.2 μ s with an error rate of 18.4%. For thinning parameter $k' > 9$, the methods lie on the ratio curve of the original data (see Fig. 12).

6 Conclusions

We showed that thinning methods and query structures for k -NN are well suited to reduce memory requirements and/or classification times for generic object recognition. The experiments showed that, for optimal speed of exact queries, the bucket size of the

kd-tree is important and independent of k . For ANN, a small bucket size and a large error bound ϵ yield the fastest queries. Furthermore, we developed k' -NN extensions of CNN, RNN and Baram and showed that they allow to tune the trade-off between data reduction and classifier degradation. As expected, the classical versions of the algorithms ($k' = 1$) yield maximum degradation. The best trade-off between query time and error rate was reached for a combination of k' -NN Baram and ANN. Gabriel and GSASH thinning turned out not to work well on the high-dimensional ETH-80 data.

References

1. Mattern, F., Denzler, J.: Comparison of appearance based methods for generic object recognition. *Pattern Recognition and Image Analysis* **14** (2004) 255–261
2. Toussaint, G.: Geometric proximity graphs for improving nearest neighbor methods in instance-based learning and data mining. *Int. J. of Comp. Geom. & Appl.* **15** (2005) 101–150
3. Clarkson, K.: A randomized algorithm for closest-point queries. *SIAM Journal of Computing* **17** (1988) 830–847
4. Dobkin, D., Lipton, R.: Multidimensional searching problems. *SIAM Journal of Computing* **2** (1976) 181–186
5. Meisner, S.: Point location in arrangements of hyperplanes. *Information and Computation* **2** (1993) 286–303
6. Yao, A., Yao, F.: A general approach to d -dimension geometric queries. In: 17th Symposium on Theory of Computing. (1985) 163–168
7. Friedman, J., Bentley, J., Finkel, R.: An algorithm for finding best matches in logarithmic expected time. *ACM Transactions on Mathematical Software* **3** (1977) 209–226
8. Maneewongvatana, S., Mount, D.: Analysis of approximate nearest neighbor searching with clustered point sets. In: *The DIMACS Series in Discrete Mathematics and Theoretical Computer Science*. Volume 59. (2002) 105–123
9. Arya, S., Mount, D.: Approximate nearest neighbor queries in fixed dimensions. In: *Fifth Annual ACM-SIAM Symposium on Discrete Algorithms*. (1993) 271–280
10. Arya, S., Mount, D., Netanyahu, N., Silverman, R., Wu, A.: An optimal algorithm for approximate nearest neighbor searching. *Journal of the ACM* **45** (1998) 891–923
11. Hart, P.E.: The condensed nearest neighbor rule. *IEEE Transactions on Information Theory* **14** (1968) 515–516
12. Gates, W.: The reduced nearest neighbor rule. *IEEE Transactions on Information Theory* **18** (1972) 431–433
13. Baram, Y.: A geometric approach to consistent classification. *Pattern Recognition* **13** (2000) 177–184
14. Olorunleke, O.: *Decision Rules for Classification: Classifying Cars into City-Cycle Miles per Gallon Groups*. Dep. of Computer Science, University of Saskatchewan, Canada (2003)
15. Toussaint, G., Bhattacharya, B., Poulsen, R.: The application of voronoi diagrams to non-parametric decision rules. In: 16th Symp. on Comp. Science and Statistics. (1984) 97–108
16. Sánchez, J., Pla, F., Ferri, F.: Prototype selection for the nearest neighbor rule through proximity graphs. *Pattern Recognition Letters* **18** (1997) 507–513
17. Mukherjee, K.: *Application of the Gabriel graph to instance based learning algorithms*. PhD thesis, Simon Fraser University (2004)
18. Leibe, B., Schiele, B.: Analyzing Appearance and Contour Based Methods for Object Categorization. In: *Int. Conf. on Comp. Vision and Pattern Recog.* Volume 2. (2003) 409–415

Non Orthogonal Component Analysis: Application to Anomaly Detection

Jean-Michel Gaucel, Mireille Guillaume, and Salah Bourennane

Institut Fresnel / CNRS UMR 6133 - EGIM,
D.U. Saint Jérôme F-13397 Marseille Cedex 20 France

Abstract. Independent Component Analysis (ICA) has shown success in blind source separation. Its applications to remotely sensed images have been investigated recently. In this approach, a Linear Spectral Mixture (LSM) model is used to characterize spectral data. This model and the associated linear unmixing algorithms are based on the assumption that the spectrum for a given pixel in an image is a linear combination of the end-member spectra. The assumption that the abundances are mutually statistically independent random sources requires the separating matrix to be unitary. This paper considers a new approach, the Non Orthogonal Component Analysis (NOCA), which enables to relax this assumption. The experimental results demonstrate that the proposed NOCA provides a more effective technique for anomaly detection in hyperspectral imagery than the ICA approach. In particular, we highlight the fact that the difference between the performances of the two approaches increases when the number of bands decreases.

1 Introduction

Over the past years, linear spectral mixture analysis has been widely used for hyperspectral image analysis such as detection and classification [1, 2, 3]. It assumes that an image pixel is linearly mixed by materials with relative abundance fractions present in the image. The observations are modeled as convex combinations of constituent spectra, known as end-members. Two approaches are possible; the first one consists in looking for the end-members. Several different procedures have been developed to automatically find them from the hyperspectral data. Then the abundance estimated are obtained as the solution of a constrained least squares problem. The second approach considers the abundance as a random signal source. This one allows to capture their spectral variability more effectively in a stochastic manner. We are in the framework of blind source separation.

Blind source separation (BSS), which consists in recovering original signals from their mixtures when the mixing process is unknown, has been a widely studied problem in signal processing for the last two decades (for a review see [4]). Independent component analysis (ICA), a statistical method for signal separation [5, 6, 7] is also a well-known issue in the community. Its aim is to transform the mixed random signals into source signals or components which are as mutually independent as possible.

In remotely sensed imagery, the number of target pixels of interest, such as small man-made targets, anomalies, or rare minerals, is generally small compared to the image background. The unsupervised methods are known as anomaly detection. From this point of view, an interesting structure of an image scene is the one resulting from a small number of target pixels in a large area of unknown background. As a consequence, these target pixels are main causes of outliers of distributions which can be detected by higher order statistics such as skewness or kurtosis. Conversely, the background which is composed of a great number of pixels can be assumed to be a Gaussian distribution while the target pixels of interest can be viewed as non-Gaussian signal sources that create ripples in the Gaussian tails. In this case, target pixels of interest can be separated by the ICA, as we desire.

The first step of many variants of the ICA algorithms consists in removing the sample mean and a whitening. Then the ICA problem can be formulated as the one to find the separating matrix \mathbf{W} derived from the ICA that separates the original signals from the mixture. By assumption \mathbf{W} is assumed to be a unitary matrix in the whitening space. In remotely sensed imagery context, \mathbf{W} is the estimate of the whitening end-members. Because of this constraint, some targets can hide other ones if the targets are correlated. We propose a new approach which relaxes this constraint. Called Non Orthogonal Component Analysis (NOCA), this approach estimates each line of \mathbf{W} iteratively after the removal of the contribution of the targets already detected.

The paper is organized as follows. Section 2 describes the Linear Mixture Model. Section 3 develops the ICA approach for hyperspectral image analysis, interprets the thresholding of the abundance map, and explains the limits of the ICA approach. Section 4 presents the Non Orthogonal Component Analysis (NOCA). Section 5 presents experimental results. Finally, section 6 includes some concluding remarks.

2 Linear Mixture Model

Linear mixture models have been extensively used to characterize spectral data, see [1, 2, 3]. These models and the associated linear unmixing algorithms are based on the assumption that the spectrum for a given pixel in an image is a linear combination of the end-member spectra.

Let \mathbf{r} be a $L \times 1$ column pixel vector in a multispectral or hyperspectral image. Let \mathbf{M} be a $L \times p$ end-member signature matrix, denoted by $[\mathbf{m}_1 \mathbf{m}_2 \dots \mathbf{m}_p]$ where \mathbf{m}_j is a $L \times 1$ column vector represented by the j th end-member signature and p is the total number of end-members in the image. Let $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_p)^T$ be a $p \times 1$ abundance column vector associated with \mathbf{r} , where α_j denotes the abundance fraction of the j th end-member signature present in the pixel vector \mathbf{r} . The pixel vector \mathbf{r} can be represented by a linear regression model as follows:

$$\mathbf{r} = \mathbf{M}\alpha + \mathbf{n}, \quad (1)$$

where \mathbf{n} is a noise that can be interpreted as measurement error, noise, or model error.

A spectral linear unmixing method estimates the unknown abundance fractions $\alpha_1, \alpha_2, \dots, \alpha_p$ via an inverse of a linear mixture model described in (1).

One requirement is that the target signature matrix \mathbf{M} must be known *a priori*. Many approaches have been proposed in the past to obtain \mathbf{M} directly from the image data in an unsupervised fashion, such as [8, 9]. Given the end-members, the abundance estimates are obtained, for example, as the solution of a constrained least squares problem.

In an other way, we can assume that the p abundance fractions $\alpha_1, \alpha_2, \dots, \alpha_p$ are unknown random quantities specified by random signal sources rather than unknown deterministic quantities, as assumed in model (1). The set of the pixel vectors of the image $\mathbf{R} = [\mathbf{r}_1, \mathbf{r}_2, \dots, \mathbf{r}_N]$ represents the set of the realizations of \mathbf{r} . In this case, we can consider this model as an inverse problem of blind source separation. This approach is developed in the following section.

3 Independent Component Analysis

3.1 Model

The independent component analysis has widely proved itself as far as blind source separation is concerned. To remain in the limits of its application, we need to make the three following additional assumptions on the random abundance vector $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_p)^T$:

- The p target signatures $\mathbf{m}_1 \mathbf{m}_2 \dots \mathbf{m}_p$ in \mathbf{M} must be spectrally distinct.
- The p abundance fractions $\alpha_1, \alpha_2, \dots, \alpha_p$ are mutually statistically independent random sources.
- Each of the p abundance fractions $\alpha_1, \alpha_2, \dots, \alpha_p$ must be a zero-mean random source and at most one source is Gaussian.

Except for these three assumptions, no prior knowledge is assumed about the model (1).

In order to implement the ICA using model (1), the mixing matrix used in the blind source separation is replaced with the target signature matrix \mathbf{M} and the unknown signal sources to be separated with the target random abundance fractions are denoted by $\alpha_1, \alpha_2, \dots, \alpha_p$. With this interpretation, the ICA finds a $p \times L$ separating matrix \mathbf{W} and applies it to an image pixel to unmix the $\alpha_1, \alpha_2, \dots, \alpha_p$. More specifically, the ICA solves an inverse problem of model (1) for a $p \times L$ separating matrix \mathbf{W} via the following equation:

$$\hat{\alpha}(\mathbf{r}) = \mathbf{W}\mathbf{r}, \quad (2)$$

where $\hat{\alpha}(\mathbf{r}) = (\hat{\alpha}_1(\mathbf{r}), \hat{\alpha}_2(\mathbf{r}), \dots, \hat{\alpha}_p(\mathbf{r}))^T$ is the estimate of abundance fractional vector $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_p)^T$ based on \mathbf{r} .

The separation is only unique up to a scaling and ordering of the components α_k . Unless we are interested in quantification, the order and the true abundance fractions are generally not crucial in target detection and classification. In this

case, we can normalize each abundance source to unit variance so that the covariance matrix of the abundance sources becomes the identity matrix. This can be simply done by a sphering (whitening) process.

In order to use the ICA, a criterion is required to measure the statistical independence among the estimated abundance fractions $\hat{\alpha}_1(\mathbf{r}), \hat{\alpha}_2(\mathbf{r}), \dots, \hat{\alpha}_p$. According to information theory [10], relative entropy or Kullback-Leibler information distance function is an appropriated measure. Nevertheless, instead of minimizing the statistical dependence between α_i , Comon suggested to maximize the nongaussianity of the estimated $\hat{\alpha}_i$ distribution. He proposed to maximize the higher order statistics of the data, called contrast function of \mathbf{W} , denoted by $\psi(\mathbf{W})$ which is defined in [5] by

$$\psi(\mathbf{W}) = \sum_{i=1}^p K_{ii\dots i}^2 \tag{3}$$

where $K_{ii\dots i}$ are marginal standardized cumulants of order $r \geq 2$ of the $\hat{\alpha}_i$ distribution.

An ICA algorithm can be summarized by the following procedure:

- Centering: remove mean of \mathbf{r} , $\tilde{\mathbf{r}} = \mathbf{r} - E[\mathbf{r}]$
- Whitening: whitened observed data \mathbf{r} through eigenvalue decomposition of the covariance matrix of measurement,

$$\tilde{\mathbf{r}} = \mathbf{D}^{-1/2} \mathbf{U}^T \mathbf{r}, \tag{4}$$

in which \mathbf{U} is the orthogonal matrix of eigenvectors of $E[\tilde{\mathbf{r}}\tilde{\mathbf{r}}^T]$ and \mathbf{D} is the diagonal matrix of its eigenvalues.

- Estimating independent components \mathbf{M} by finding the projection matrix \mathbf{W} solution of the following optimization problem:

$$\begin{aligned} & \text{maximize} && \psi(\mathbf{W}) \\ & \text{subject to} && E[\hat{\alpha}(\mathbf{r})\hat{\alpha}(\mathbf{r})^T] = \mathbf{Id} \end{aligned} \tag{5}$$

where \mathbf{Id} is the $p \times p$ identity matrix.

The estimate of the independent component is given by

$$\hat{\alpha}(\mathbf{r}) = \mathbf{W}\tilde{\mathbf{r}}, \tag{6}$$

and the p abundance detectors are

$$\hat{\alpha}_k(\mathbf{r}) = \mathbf{w}_k \tilde{\mathbf{r}}, \quad k = 1 \dots p, \tag{7}$$

where $\mathbf{W} = [\mathbf{w}_1^T, \mathbf{w}_2^T, \dots, \mathbf{w}_p^T]^T$.

3.2 ICA and Detection

In this part, we will try to make the link between the independent component analysis and the detection theory.

Let

$$\mathbf{S}^T = \mathbf{W}\mathbf{D}^{1/2}\mathbf{U}, \tag{8}$$

then as \mathbf{U} is unitary

$$\mathbf{W} = \mathbf{S}^T\mathbf{U}^T\mathbf{D}^{-1/2}. \tag{9}$$

With the equation (4) and the equation (9), the equation (6) becomes

$$\hat{\alpha}(\mathbf{r}) = \mathbf{S}^T\mathbf{U}^T\mathbf{D}^{-1/2}\mathbf{D}^{-1/2}\mathbf{U}^T\bar{\mathbf{r}} = \mathbf{S}^T E[\mathbf{r}\mathbf{r}^T]^{-1}\bar{\mathbf{r}}. \tag{10}$$

Let us note $\mathbf{S} = [\mathbf{s}_1, \mathbf{s}_2 \dots \mathbf{s}_P]$. In these conditions we can write:

$$\hat{\alpha}_k(\mathbf{r}) = \mathbf{s}_k^T E[\mathbf{r}\mathbf{r}^T]^{-1}(\mathbf{r} - E[\mathbf{r}]). \tag{11}$$

If we notice that:

$$\begin{aligned} \mathbf{S}^T E[\mathbf{r}\mathbf{r}^T]^{-1}\mathbf{S} &= \mathbf{W}\mathbf{D}^{1/2}\mathbf{U}E[\mathbf{r}\mathbf{r}^T]^{-1}\mathbf{U}^T\mathbf{D}^{1/2}\mathbf{W}^T \\ &= \mathbf{W}\mathbf{W}^T \\ &= \mathbf{Id}, \end{aligned} \tag{12}$$

then $\mathbf{s}_k^T E[\mathbf{r}\mathbf{r}^T]^{-1}\mathbf{s}_k = 1$ and the equation (11) can be written as

$$\hat{\alpha}_k(\mathbf{r}) = \frac{\mathbf{s}_k^T E[\mathbf{r}\mathbf{r}^T]^{-1}(\mathbf{r} - E[\mathbf{r}])}{\mathbf{s}_k^T E[\mathbf{r}\mathbf{r}^T]^{-1}\mathbf{s}_k}, \tag{13}$$

which is the expression of the Adaptive Matched Filter (AMF) for the target signature \mathbf{s}_k . The AMF filter [11] is the most popular Constant False Alarm Rate (CFAR) of the hyperspectral imaging target detection algorithms.

This remark enables us to make two important conclusions for the following. First of all we can interpret the ICA as the research of the spectral signatures which give AMF detection maps with the most distant histogram from a gaussian distribution. Given that the end-members the less present in the scene give the detection maps whose histogram is the less gaussian, we easily understand the particular interest of this method in anomaly detection.

The second conclusion is very important too. It comes from the fact that the AMF filter is a CFAR filter. In [12], Chang gave a method to determine a threshold for each abundance map from a rejection rate. However the value of this rejection rate is arbitrary. Nothing can justify that each abundance map has the same rejection rate. The property of the CFAR will enable us to threshold each abundance map independently according to a defined probability of false alarm (PFA).

3.3 Limits of the Model

As seen in the equation (12), the independent component analysis requires that the spectra of the estimated targets are orthogonal in the whitening space. However if we consider pure spectra, their mutual correlation becomes sometimes important and it makes the application of a separation method based on the mutual independence of the sources fruitless. Two correlated targets will thus

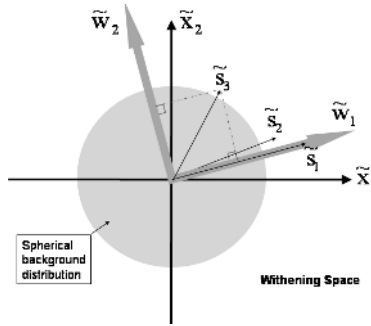


Fig. 1. Illustration of the limits of the ICA model in 2D dimension

be either identified as coming from the same source and detected on the same abundance map, or, and it would be more annoying, a target will hide an other and will not enable its detection. The figure 1 illustrates simply the problem. The direction \mathbf{w}_1 enables to detect \mathbf{s}_1 optimally. Moreover it enables to detect suboptimally \mathbf{s}_2 . But neither \mathbf{w}_1 nor \mathbf{w}_2 enable to detect \mathbf{s}_3 .

This limitation is a consequence of the hypothesis of statistical independence between the p abundance fractions $\alpha_1, \alpha_2, \dots, \alpha_p$. This hypothesis is valid on the abundances but not any more on their estimates $\hat{\alpha}_1, \hat{\alpha}_2, \dots, \hat{\alpha}_p$ because the end-members spectra are correlated.

We propose to remedy to this problem with an original method.

4 Non Orthogonal Component Analysis (NOCA)

We propose an iterative method, the Non Orthogonal Component Analysis (NOCA), for blind source separation problem, which enables to relax the orthogonality assumption imposed by the ICA. The principle is illustrated in the figure 2. First, we estimate the most independent direction \mathbf{w}_1 (2-a) from the observation. Then, we threshold the abundance map and remove from the observation the pixel vectors with an estimated abundance superior to a threshold η (here \mathbf{s}_1 and \mathbf{s}_2). After, we calculate \mathbf{w}_2 in this new set of observations (2-b). As \mathbf{s}_1 and \mathbf{s}_2 have been removed, the second independent component follows the direction of \mathbf{s}_3 . So it is impossible to detect the three targets.

We note that the two estimated independent components are not orthogonal, contrary to those in the classic method. Moreover, the NOCA approach can eventually estimate a number of components greater than the number of observed sources; while this one is increased with the classic ICA.

We only need to make the following two assumptions on the random abundance vector $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_p)^T$:

- The p target signatures $\mathbf{m}_1 \mathbf{m}_2 \dots \mathbf{m}_p$ in \mathbf{M} must be spectrally distinct.
- Each of the p abundance fractions $\alpha_1, \alpha_2, \dots, \alpha_p$ must be a zero-mean random source and at most one source is Gaussian.

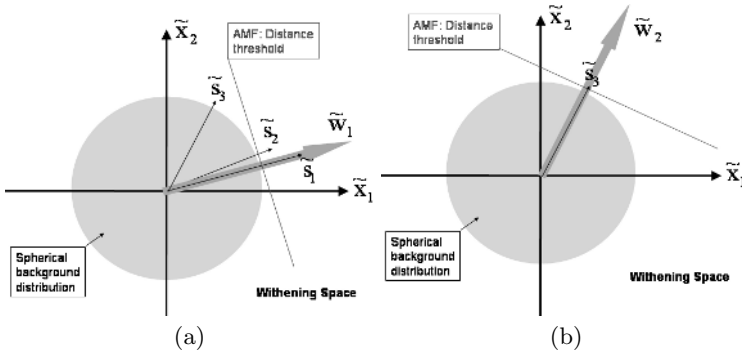


Fig. 2. The NOCA approach: (a) First component estimated; (b) Second component estimated

The assumption "The p abundance fractions $\alpha_1, \alpha_2, \dots, \alpha_p$ are mutually statistically independent random sources" begin obsolete since the estimated sources are not orthogonal.

Let $\tilde{\mathbf{R}} = [\tilde{\mathbf{r}}_1, \tilde{\mathbf{r}}_2, \dots, \tilde{\mathbf{r}}_N]$ be the set of whitened observations of the pixel vector \mathbf{r} , η a threshold, and Q the number of sources to estimate. The NOCA algorithm can be summarized by the following procedure:

- Centering: remove mean of \mathbf{r} , $\tilde{\mathbf{r}} = \mathbf{r} - E[\mathbf{r}]$
- Whitening: whitened observed data \mathbf{r} through eigenvalue decomposition of the covariance matrix of measurement,

$$\tilde{\mathbf{r}} = \mathbf{D}^{-1/2} \mathbf{U}^T \mathbf{r}, \tag{14}$$

where \mathbf{U} is the orthogonal matrix of eigenvectors of $E[\tilde{\mathbf{r}}\tilde{\mathbf{r}}^T]$ and \mathbf{D} is the diagonal matrix of its eigenvalues.

- Initialisation:

$$\begin{aligned} \tilde{\mathbf{R}}_1 &= \tilde{\mathbf{R}}, \\ k_1 &= N \end{aligned} \tag{15}$$

- For $i = 1 \dots Q$
 - Estimating the independent component α_i by finding the projection vector \mathbf{w}_i solution of the following optimization problem:

$$\text{maximize } \psi(\mathbf{w}_i) = E[\tilde{\alpha}_i^m] = E[(\mathbf{w}_i \tilde{\mathbf{r}})^m] \text{ for } m \geq 2, \tag{16}$$

for the observations $\tilde{\mathbf{R}}_i$.

- Let $\mathbf{y} = [y_1 y_2 \dots y_{k_i}] = \mathbf{w}_i \tilde{\mathbf{R}}_i$ be the estimated abundance of \mathbf{w}_i in the observations $\tilde{\mathbf{R}}_i$. Find k_{i+1} and $f : \mathbb{N} \cap [1; N] \mapsto \mathbb{N} \cap [1; N]$ bijective, so that:

$$\begin{aligned} y_{f(j)} &< \eta \text{ for } j = 1..k_{i+1} \\ y_{f(j)} &> \eta \text{ for } j = k_{i+1} + 1..k_i \end{aligned} \tag{17}$$

- Update:

$$\tilde{\mathbf{R}}_{i+1} = [\tilde{\mathbf{r}}_{f(1)} \tilde{\mathbf{r}}_{f(2)} \dots \tilde{\mathbf{r}}_{f(k_{i+1})}]. \tag{18}$$

- form $\mathbf{W} = [\mathbf{w}_1^T, \mathbf{w}_2^T, \dots, \mathbf{w}_p^T]^T$.

The estimate of the independent component is given by

$$\hat{\alpha}(\mathbf{r}) = \mathbf{W}\tilde{\mathbf{r}}. \quad (19)$$

5 Experiments: Anomaly Detection

In this section, we will compare the ICA and NOCA approaches within the framework of anomaly detection in hyperspectral imagery. Several configurations will be envisaged in order to highlight the performances of each of the two methods for this application.

There are two ways to calculate the detection probabilities. The first considers that a target is detected if one of its pixel is detected (Noted Pd_1). The second considers the set of pixel vectors composing a target as so many independent targets (Noted Pd_2)

For the ICA, we have chosen to use the FastICA, a fixed-point algorithm first proposed by Hyvrinen and Oja [6, 7]. It is one of the most widely used algorithm for the linear mixing model. In the same way, we use a fixed-point method for the NOCA to maximize our contrast function.

The contrast function used will be the same for the two methods: the kurtosis. It is used in many approaches [13, 14].

5.1 Data

A high spatial resolution hyperspectral digital imagery collection experiment (HYDICE) scene considered in most of research tests was used for experiments. The HYDICE image shown in figure 3(a) has a size of 243×113 with 10 nm spectral resolution and 1.5 m spatial resolution. The low signal / high noise bands (bands 1-3 and bands 202-210) and water vapor absorption bands (bands 101-112 and bands 137-153) have been removed. It results in a total of 169 bands. There are 30 target panels located on the field, and they are arranged in a 10×3 matrix. The figure 3(b) shows the ground truth map of 3(a) and provides the precise spatial locations of these 30 panels. The sizes of the panels in the first, second and third columns are $3m \times 3m$, $2m \times 2m$, and $1m \times 1m$, respectively. Spectra of the ten targets are represented in figure 3(c).

5.2 Influence of the Spectral Domain

We have a 148 bands image. Firstly, we will vary the number of bands. We will arbitrarily go from the band 1 to the band X , by step of 10. The results are presented in fig. 4 for a probability of false alarm equal to 5.10^{-4} .

First, we notice that the results are really good with the two methods for the total number of bands but also for fewer bands (110). Then, their performances

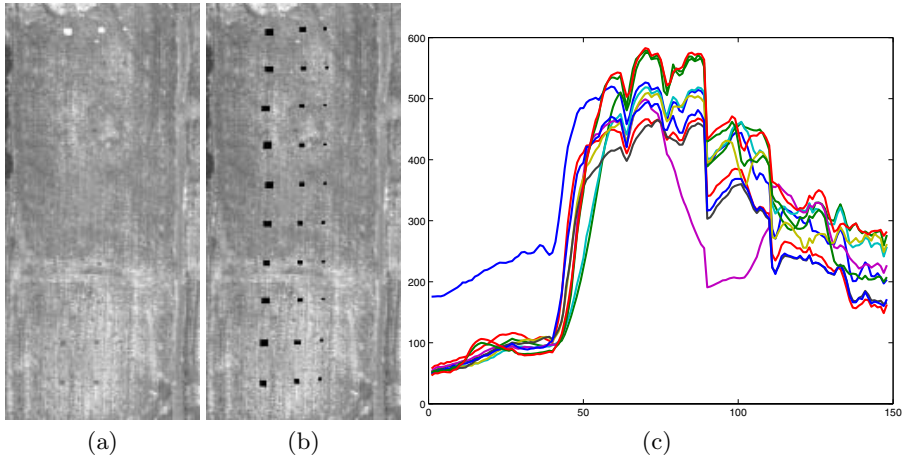


Fig. 3. (a) A 30-panels HYDICE image scene. (b) Ground truth map of target in fig. 3(a). (c) Spectra of 10 targets.

decrease but NOCA is less affected by the reduction of the number of bands. The difference of performance between the two approaches increases up to 20% for 50 bands. Finally, their performances for 40 bands are similar because only one abundance map is conserved in this configuration, there are no difference left between the two methods.

Then, we will consider groups of 40 adjacent bands and assess the performances of the two algorithms on them. The figure 5 shows the results for a probability of false alarm equal to 5.10^{-4} . We note again that the NOCA method performs better than the ICA one. Even if it is not really notable, there is always an advantage with NOCA rather than with ICA.

We notice that the smaller the number of band is, the more advantageous the NOCA approach is. It can be explained by the fact that when the dimension reduces, the correlation between the targets increases. These first results highlight the fact that the NOCA approach is adapted to anomaly detection in hyperspectral imagery. The relax of the orthogonality constraint allows to enhance the detection performances.

5.3 From Multispectral to Hyperspectral

Still in order to compare the two methods but also to assess the advantage of the hyperspectral on the multispectral in this particular context, we have under-sampled the spectral bands. The figure 6 shows the results for a probability of false alarm equal to 5.10^{-4} .

For a ratio of under-sampling equal to 1, we observe the same performances than the previous obtained for 148 bands. Then the NOCA method performs better than the ICA one, in particular when the number of bands is very reduced. This results are in keeping with the previous ones.

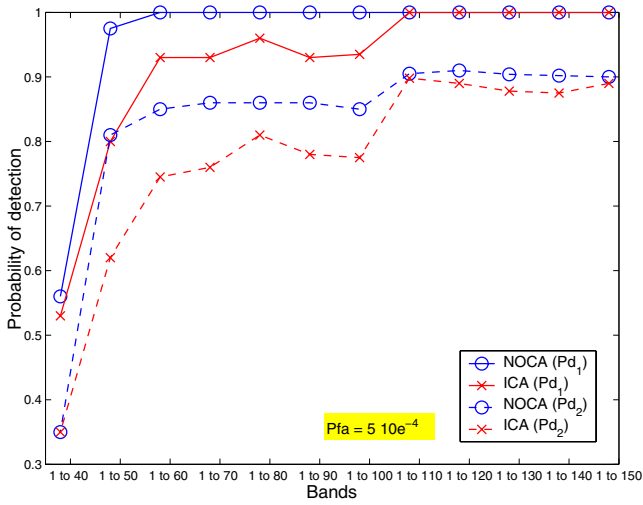


Fig. 4. Performances of ICA and NOCA algorithms according to the number of spectral bands

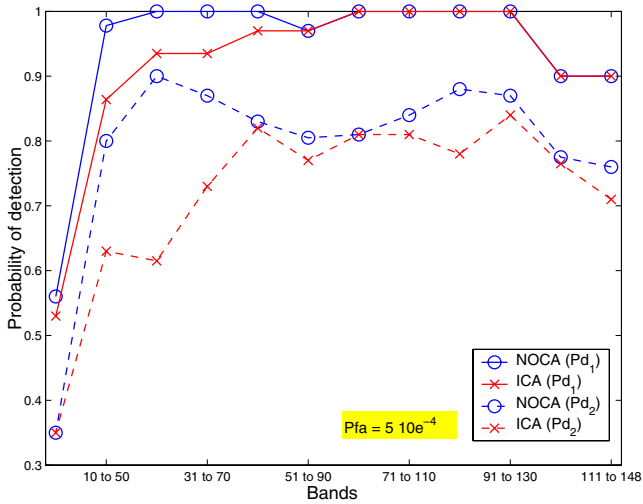


Fig. 5. Performances of ICA and NOCA algorithms according to the spectral domain with 40 spectral bands

Moreover, we notice that 9 bands enable to have performances equivalent to those of 148 bands. This result is important because it is not easy to detect all targets with 148 bands. For example the RX detector fails to do that, see [15]. We can here conclude that the advantage of hyperspectral on multispectral is not obvious on these data for this application.

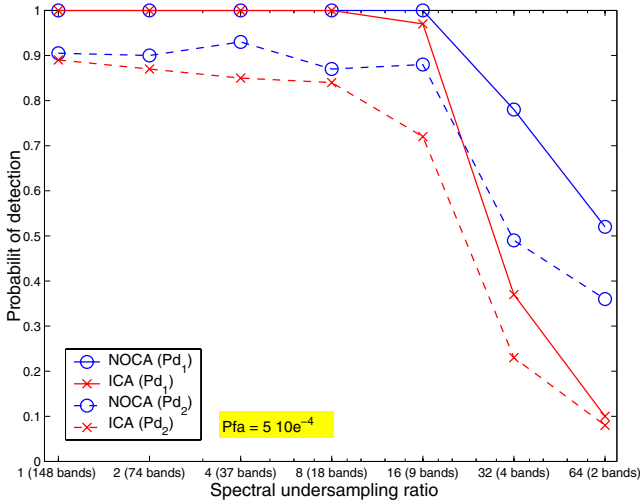


Fig. 6. Performances of ICA and NOCA from hyperspectral to multispectral

6 Conclusion

This paper has presented a new approach, the NOCA, for blind source separation problem, which relaxes the orthogonality assumption imposed by the ICA. Our presented experimental results demonstrate that the proposed NOCA provides a more effective technique for anomaly detection in hyperspectral imagery than the ICA approach. Indeed the relax of the orthogonality constraint has enabled to enhance the detection performances. We have also shown that the advantage of hyperspectral on multispectral is not obvious on the test data for this application. In the aim of completing this work, we plan to do a more theoretical study with classical blind source separation cases. Then we plan to test our new approach on unsupervised classification in hyperspectral imagery where the ICA approach has proved itself efficient.

References

- [1] J. Adams, M. Smith, and A. Gillespie, "Image spectroscopy: Interpretation based on spectral mixture analysis," in *Remote Geochemical Analysis: Elemental and Mineralogical Composition*, C. Pieters and P. Englert, Eds. Cambridge Univ. Press U.K., 1993, pp. 145–166.
- [2] M. Smith, J. Adams, and D. Sabol, "Spectral mixture analysis - new strategies for the analysis of multispectral data," in *Image Spectroscopy - A Tool for Environmental Observations*, J. Hill and J. Mergier, Eds. Brussels and Luxembourg, Belgium: ECSC, EEC, EAEC, 1994, pp. 125–143.
- [3] D. Manolakis and G. Shaw, "Detection algorithms for hyperspectral imaging applications," *IEEE Signal Proc. Mag.*, pp. 29–43, Jan. 2002.

- [4] S.-I. Amari and A. Cichocki, in *Adaptive Blind Signal and Image Processing*, J. Wiley and Sons, Eds. New York: Wiley, 2002.
- [5] P. Comon, "Independent component analysis, a new concept ?" *IEEE Trans. Signal Processing*, vol. 36, pp. 287–314, 1994.
- [6] A. Hyvrinen and E. Oja, "A fast fixed-point algorithm for independent component analysis," *Neural Comput.*, vol. 9, pp. 1483–92, 1997.
- [7] A. Hyvrinen, "Fast and robust fixed-point algorithms for independent sources: A deflation approach," *Signal Process.*, vol. 45, pp. 59–83, 1995.
- [8] C.-I. Chang and D. Heinz, "Subpixel spectral detection for remotely sensed images," *IEEE Trans. Geosci. Remote Sensing*, vol. 38, pp. 1144–59, May. 2000.
- [9] H. Ren and C.-I. Chang, "A generalized orthogonal subspace projection approach to unsupervised multispectral image classification," *IEEE Trans. Geosci. Remote Sensing*, vol. 38, pp. 2515–28, Nov. 2000.
- [10] T. Cover and J. Thomas, "Elements of information theory," in *New York: Wiley*, 1991.
- [11] F. C. Robey, D. R. Fuhrmann, E. J. Kelly, and R. Nitzberg, "A cfar adaptive matched filter detector," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 28, no. 1, pp. 208–218, Jan. 1992.
- [12] C.-I. Chang, S.-S. Chiang, J. A. Smith, and I. W. Ginsberg, "Linear spectral random mixture analysis for hyperspectral imagery," *IEEE Trans. Geosci. Remote Sensing*, vol. 40, no. 2, pp. 375–392, Feb. 2002.
- [13] C. Papadias, "Globally convergent blind source separation based on a multiuser kurtosis maximization criterion," *IEEE Trans. Signal Processing*, vol. 48, no. 10, pp. 3508–19, Dec. 2000.
- [14] J.-F. Cardoso and B. Laheld, "Equivariant adaptive source separation," *IEEE Trans. Signal Processing*, vol. 44, no. 12, pp. 3017–30, Dec. 1996.
- [15] C. Chang, in *Hyperspectral Imaging*, K. Academic, Ed. Plenum, 2003.

A Rough Set Approach to Video Genre Classification

Wengang Cheng, Chang'an Liu, and Xingbo Wang

Department of Computer Science, North China Electric Power Univ.
102206 Beijing, China
wgcheng@ncepu.edu.cn

Abstract. Video classification provides an efficient way to manage and utilize the video data. Existing works on this topic fall into this category: enlarging the feature set until the classification is reliable enough. However, some features may be redundant or irrelevant. In this paper, we address the problem of choosing efficient feature set in video genre classification to achieve acceptable classification results but relieve computation burden significantly. A rough set approach is proposed. In comparison with existing works and the decision tree method, experimental results verify the efficiency of the proposed approach.

1 Introduction

Automatic video content classification is a necessary way to efficient access, understanding and retrieval of videos. The semantic content, i.e. the story line told by the video, can be split into genre, events and objects. Correspondingly, video content classification can be carried out at above three different levels. Specially, a genre is simply a categorization of certain types of art based upon their style, form, or content. The genre of a video is the broad class to which it may belong, e.g. news, sports and cartoon. In this paper, we address the problem of video classification at the highest level: Genre.

Different methods have been introduced to categorize the video into predefined genres automatically or semi-automatically. Fischer et al. made the first attempt and a three-step approach was presented [1]. Firstly, basic acoustic and visual statistics were collected. Secondly, style attributes were derived from low-level statistics. Finally, the distributions of these style attributes are used to discriminate genres. Based on the human perception and visual characteristics, Truong et al. [2] analyzed a set of computational features including editing effects, motion and color for genre labeling and decision tree was used to verify these features. Liu et al. [3,4] investigated a range of statistical time and frequency features extracted from the acoustic signal and they used Neural Network and Hidden Markov Models (HMM) to accomplish the classification, respectively. Statistics based method, e.g. Gaussian Mixture Models (GMM), was introduced into video genre modeling in [5], focusing on direct analysis of the relationship between the probabilistic distributions of low-level audio and/or video features and the associated genre identity. In [6], motion pattern was represented by motion texture and Support Vector Machine (SVM) was used as the classifier to map video shots to semantic categories. Fifteen MPEG-7 audio-visual descriptors were fed into the decision tree-based classifier to recognize the video genres [7]. Guided by the

film grammar, [8] proposes a method to classify movies based on audio-visual cues present in the previews.

We summarize the related works from two viewpoints: classifiers and features. A key issue is classifier selection. Rule-based, decision tree-based, HMM, GMM, SVM and some other classifiers have been employed. Among them, the decision tree based method is the most used one, for it can get reliable and stable results, and the deduced rules from it are intelligible. Feature set is the other key issue. As often one single feature is not sufficient to obtain satisfactory results, most of the existing methods resort to combined features, which consist of audio and visual features, low-level and mid-level features, local and holistic features. For example, a lot of visual features used in [2,8], plenty of audio features in [3,4] and even a good few audio-visual features in [1,5,7]. In fact, these works fall into this category: enlarging the feature set until the classification is reliable enough. However, some features are *irrelevant* or *less relevant* to classification and some features are *equivalent*, which sometimes lead to redundancy. The redundant features will enlarge the complexity of rule generation and weaken the quality of the deduced rules. In addition, some automatic video processing techniques (e.g. object detection) are still immature nowadays, especially the gap between low-level features and high-level semantics, which all result in inaccurate features and the uncertainty of classification results. Especially, due to the characteristics of video data, computational analysis of these features is often time-consuming. Therefore, it is necessary to select an appropriate feature set to achieve an acceptable classification result but relieve computation burden significantly. Unfortunately, there is still no published works tackling this issue.

Rough set theory is an attempt to dispose a formal framework for the automated transformation of data into knowledge. Pawlak[9] points out that one of the most important and fundamental notions to the rough sets philosophy is the need to discover redundancy and dependencies between features. A rough set approach to video feature selection and genre classification is proposed in this paper. We use the concept *reduct* and *core* for feature selection, and the *rule set* derived from the *reduct* is used to label the genres of video clips. When an unlabeled video clip comes into system, the classifier can decide its class with fewer features.

The rest of paper is organized as follows. Section 2 provides the rough set model of video classification, the original feature set and the feature selection method, Experimental results are presented in Section 3 and Section 4 concludes the paper.

2 Rough Set, Feature Set and Feature Selection

2.1 Fundamentals of Rough Set Theory

In rough set theory, an information table is defined as a tuple $S=(U, A)$, where U and A are two finite, non-empty sets, U the universe of primitive objects and A the set of attributes [10]. We may partition the attribute set A into two subsets C and D , called condition and decision attributes, respectively. An equivalence relation, *indiscernibility relation*, is associated with every subset of attributes $P \subset A$. This relation is defined as:

$$IND(P) = \{(x, y) \in U \times U : \forall a \in P, a(x) = a(y)\} . \quad (1)$$

where $a(x)$ denotes the value of feature a of object x . If $(x, y) \in IND(P)$, x and y are said to be indiscernible with respect to P . The family of all equivalence classes of $IND(P)$ (Partition of U determined by P) is denoted by $U/IND(P)$. Each element in $U/IND(P)$ is a set of indiscernible objects with respect to P . Equivalence classes $U/IND(C)$ and $U/IND(D)$ are called condition and decision classes.

For any concept $X \subseteq U$ and attribute subset $R \subseteq U$, X could be approximated by the R -lower approximation and R -upper approximation using the knowledge of R . The lower approximation of X is the set of objects of U that are surely in X , while the upper approximation of X is the set of objects of U that are possibly in X . They are defined as:

$$\underline{R}(X) = \bigcup \{E \in U / IND(R) : E \subseteq X\} . \quad (2)$$

$$\overline{R}(X) = \bigcup \{E \in U / IND(R) : E \cap X \neq \emptyset\} . \quad (3)$$

The boundary region, $BND_R(X)$, is the difference of $\overline{R}(X)$ and $\underline{R}(X)$. If $BND_R(X)$ is not empty, X is a rough set with respect to R . The positive region of decision classes $U/IND(D)$ with respect to condition attributes C is denoted by $POS_C(D) = \bigcup \underline{R}(X)$. It is a set of objects of U that can be classified with certainty to classes $U/IND(D)$ employing attributes of C . A subset $R \subseteq C$ is said to be a D -reduct of C if $POS_R(D) = POS_C(D)$ and there is no $R' \subseteq R$ such that $POS_{R'}(D) = POS_C(D)$. In other words, a *reduct* is the minimal set of attributes preserving the positive region.

In this paper, the video classification system is the decision table S . Evidently, $IND(P)$ is a equivalence relation, while P can be regarded as a name of the knowledge represented by $IND(P)$. Hereby, S serves as a knowledge base system. Therefore, an unlabeled video clip can be classified using the rules corresponding to the $IND(P)$ of S . Another direct application of the rough set theory, shown in Section 2.3, is feature selection based on the method of finding the *reduct*. Although *reduct* computation is considered the bottleneck in the rough set methodology, many good heuristics can accomplish it in a reasonable amount of time.

2.2 Feature Extraction

An appropriate feature set is crucial to video classification. It should be not only a set of content descriptors, but also a basis for comparison between different genres. Different with earlier works, we do not target at exploring new descriptors, but pursuing a powerful feature set to achieve an acceptable classification. Hence, we collect a set of audio-visual features, which have shown their discriminabilities in the related works but some may in different forms. Here, we just list them with brief explanation. Four types of features are used:

Editing features. Editing style in a video provides good indication about its genre. There are many shots of smaller duration in fast paced videos (action movies/MTVs), while shot duration are often large in slow paced videos (documentaries/dramas). News, for example, has large number of abrupt transitions, while there are more gradual transitions in the commercials. The following four features are computed: F_1 (average shot length); F_2, F_3, F_4 (the percent of each type of shot transition, three most used transition types, i.e. cut, fade, dissolve, are considered).

Motion Feature. Different genres have different motion intensity and show different motion patterns. For instance, sports video has larger motion intensity than news. The MPEG-7 motion activity descriptor captures the intuitive notion of “intensity of action” or “pace of action” in a video segment. F_5 (the average intensity of motion activity) is calculated as:

$$IA = \left(\sum_{j=1}^{N_p} IAF_j \right) / N_p . \quad (4)$$

Color Features. There are also distinctions in color characters between different genres. Due to the special making rules, the brightness and saturation of cartoons are much larger than those of other genres. MTVs and commercials often have quick change of lighting, so they have much color variance. F_6 (average brightness), F_7 (average saturation) and F_8 (percent of color variance) are used as color features.

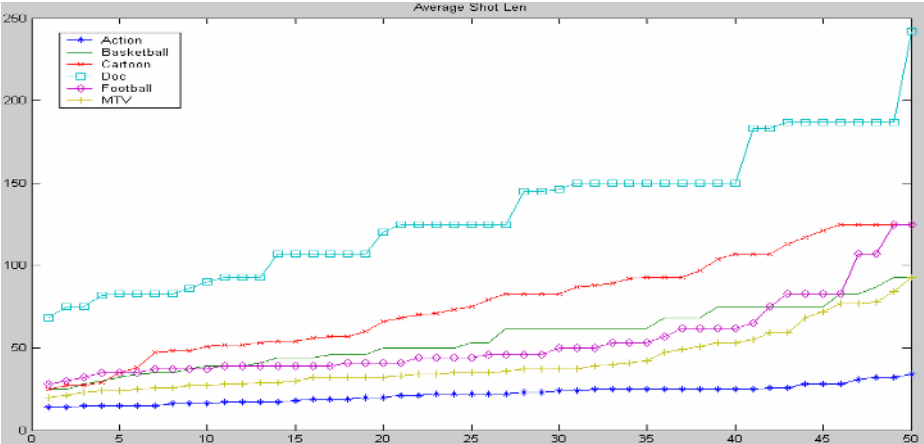
Audio features. By audio cues, man can distinguish different genres easily even without watching the moving pictures. We collect three mid-level features F_9 (silence ratio), F_{10} (noise ratio) and F_{11} (background noise level). For the following nine frame-level features: pause rate, number of low energy windows, sum of scalefactors as loudness approximation, spectral centroid, short time energy, short time band energy, number of low energy windows, short time magnitude and spectral flux, we calculate their mean and variance as the holistic features ($F_{12} \sim F_{29}$). All the audio features are extracted using the Maaate toolkit [11].

Fig.1 illustrates several distributions of feature values in our experiments (for 50 video samples randomly selected from each genre) after sorting in ascending order. It shows that these features have certain discriminabilities.

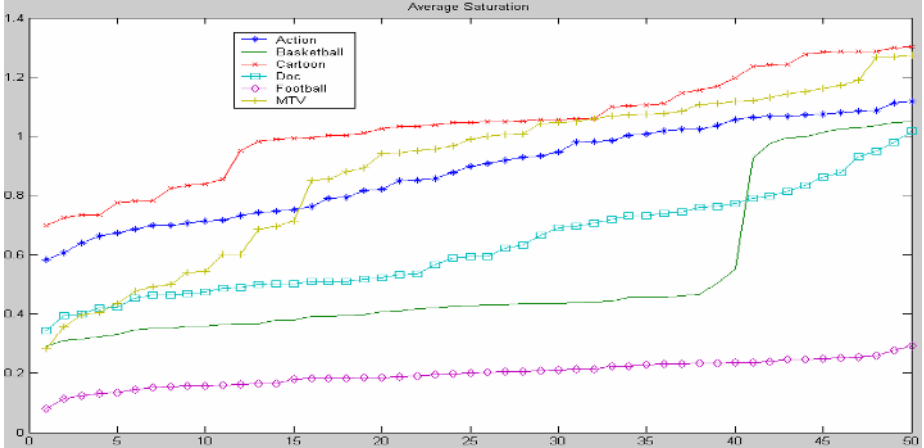
2.3 Feature Selection

Features selection is a process to find the optimal subset of the original feature set that satisfies certain criteria. The original feature set is the set C in model of video classification system S . Let R be a set of selected features, P a set of unselected features, we get $C = R \cup P$.

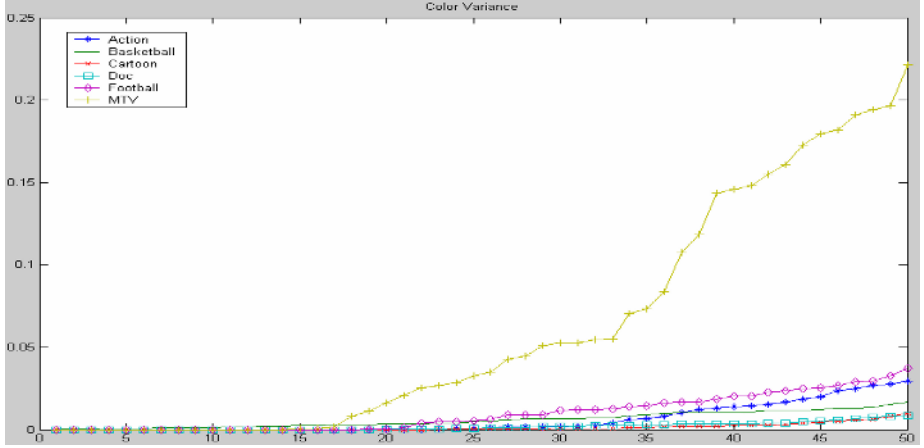
Feature selection can be viewed as a search problem according to some evaluation criterion. The optimal feature subset is the one that maximizes the value of evaluation measure. Compared with exhaustive search and random search, the heuristic search



(a)



(b)



(c)

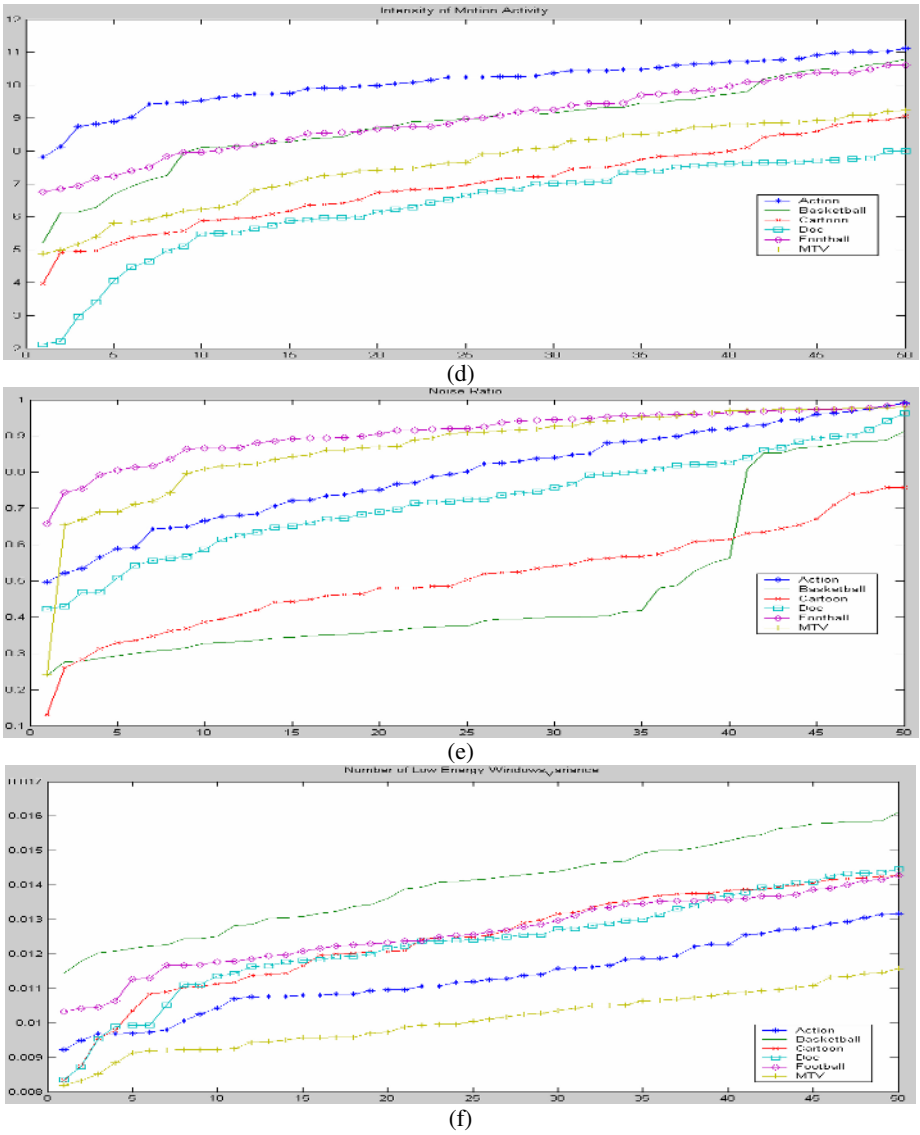


Fig. 1. Distribution of several feature values: (a) average shot length; (b) average saturation; (c) color variance; (d) intensity of motion activity; (e) noise ratio; (f) variance of num of low energy windows

can achieve good tradeoff between computational complexity and search result. Therefore, it is the most commonly used method. In the heuristic search, a heuristic function is employed to guide the search and the search is performed towards the direction that maximizes the function value. In this paper, the solution based on rough

set theory also follows this strategy. According to the basic issues of heuristic feature selection, we describe our method as follows.

Starting Point. Rough set theory defines the *core* of an information system to be the set of indispensable features. Removal of any attribute from the *core* set changes the positive region with respect to the label. This fact can be interpreted as a similarity between the *core* set and the notion of strong relevance introduced by John et al. [12]. In other words, the *core* is the set of strong relevant features. We use the *core* as the initial feature subset, i.e., $R = CORE(C)$. The *core* is computed using the method based on discernibility matrix in our experiments.

Stop criteria. In rough set based methods, the size of the positive region could be used as stop criteria. In particular, the algorithm stops when the positive region of the selected features reaches the original positive region, i.e., $POS_R(D) = POS_C(D)$, which also manifests that feature selection is a process of finding the optimal *reduct*.

Search Origination. The attributes of the *core* may be insufficient for defining all decision classes. Therefore, other attributes may be added to the *core* in order to maintain the same classification power that the one achieved with all the features. We adopt the greedy forward direction method: at each iteration, the best feature a of P is chosen and added to the R until the stop criteria is reached, in which the heuristic function is to evaluate the feature subsets and select to optimal feature a .

In Rough set, significance of a feature a , denoted as $SIG(a)$, is the increase of dependency between condition attributes and decision attribute as a result of the addition of a :

$$SIG(a) = \gamma_{(R+\{a\})}(D) - \gamma_R(D) . \quad (5)$$

where $\gamma_R(D)$ is the dependency between R and D , it reflects the importance of R in classifying the objects into the decision classes: $\gamma_R(D) = card(POS_R(D))/card(U)$. [13] adopt the Equation (5) as heuristic function for its simplicity and low complexity. However, this function only considers the dependency of the selected features, but ignores the quality of the potential rules. As the ultimate goal of feature selection is to reduce the number of features used to generate classification rules, we want get high quality rules. The quality of the rules can be evaluated by two parameters: 1) the number of instances covered by the potential rules, that is, the size of consistent instances; and 2) the number of instances covered by each rule, called support of each rule. Significance oriented methods, for example [13], only consider the first parameter. It attempts to increase faster the size of consistent instances but ignoring the support of individual rules. However, rules with very low support are usually of little use. Therefore, we adopt the method considering both parameters in [14], and the heuristic function is defined as:

$$F(R, a) = card(POS_{(R+\{a\})}(D)) \times MaxSize(POS_{R+\{a\}}(D) / IND(R + \{a\} + D)) . \quad (6)$$

where $card(POS_{(R+\{a\})}(D))$ is the size of consistent instances and $MaxSize(POS_{R+\{a\}}(D)/IND(R+\{a\}+D))$ denotes the maximal size out of indiscernibility classes included in the positive region, i.e., the support of the most significant rule, when a added. The feature selection in our work are performed as follows steps:

- (1) Initialize: Get the *core* $CORE(C)$ of video classification system, set $R = CORE(C), P = C - R$;
- (2) Remove all consistent instances: $U = U - POS_R(D)$. If $U = \emptyset$, stop and return R ;
- (3) For each $a \in P$, calculate $F(R, a)$. Choose the best feature a with largest $F(R, a)$, set $R = R \cup \{a\}, P = P - \{a\}$;
- (4) If $POS_R(D) = POS_C(D)$, return R and stop; else goto step (2).

3 Experimental Results

The block diagram of our classification system is illustrated in Fig. 2. It consists of two parts, off-line rule generation and on-line video classification. At off-line stage, the original feature set FC is formed after feature extraction for learning video clips, which then must be discretized before *reduct* computation. Once the *reduct* R has been computed, it is used to generate rules. At on-line stage, the features corresponding to the *reduct* R are firstly extracted when unlabeled video clips comes. And then, the feature set FR is discretized using the cuts set $CUTS$ produced at off-line stage. Finally, $RULES$ derived from the *reduct* R works on the discretized feature set and return the results.

3.1 Experimental Dataset

In order to perform video genre identification fast and efficiently, long videos are usually segmented into short time video clips, and some of these clips are chosen randomly as dataset. We collect video clips of six genres: action movie (AM), football (FB), basketball (BB), documentary (DOC), cartoon (CT) and MTV (MTV), 80 video clips for each genre. The duration of each video clip is 60s, which is moderate compared with that of existing works. 300 clips (6*50) are randomly chosen as learning set, the others for test.

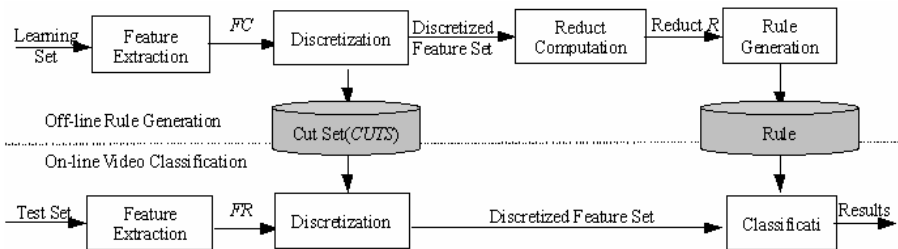


Fig. 2. Diagram of video genre classification based on rough set

3.2 Results and Evaluation

Although there are various methods for discretization, the two-step strategy (the boolean algorithm applied first and then equal frequency binning method used on the remaining data) gives best results in our experiments. Based on the methods for finding *core* and *reduct*, we get *core* as $CORE(C)=\{F_{13}\}$, and the selected feature set as $R=\{F_1, F_3, F_5, F_7, F_{13}, F_{14}, F_{16}, F_{24}, F_{26}, F_{28}\}$. The discretized features for the test set are classified using the standard voting with rule support. Some related algorithms provided by the Rosetta platform[15] are used in our experiments.

The classification results are list in Table 1. Confusion matrix M is used to summarize the results. The entries $M(i, j)$ is the number of objects classified as belonging to classes i that actually belong to class j . For each class k , the precision and recall rate are calculated as:

$$Precision(k) = M(k, k) / \sum_l M(l, k) . \quad (7)$$

$$Recall(k) = M(k, k) / \sum_l M(k, l) . \quad (8)$$

The average accuracy that measures the overall performance of classifier is:

$$AvgAccuracy = (\sum_k M(k, k)) / \sum_{i,j} M(i, j) . \quad (9)$$

Table 1. Classification results based on the proposed method

	AM	FB	BB	DOC	CT	MTV	Recall
AM	27	0	3	0	0	0	0.90
FB	0	27	3	0	0	0	0.90
BB	0	0	30	0	0	0	1.00
DOC	0	0	9	21	0	0	0.70
CT	0	0	4	1	25	0	0.83
MTV	0	0	10	0	0	20	0.67
Precision	1.00	1.00	0.51	0.95	1.00	1.00	0.83

The average accuracy of our method is 83.3%. [2] works on five genres: commercials, news, sports, MTV and cartoons. When the clip duration is 40s and 60s, the average accuracy is 80% and 83.1%, respectively. Liu et al.[4] classify basketball, football and weather forecast using HMM with five states, the average accuracy is around 84.7%. In a word, there is little difference between our results and theirs. However, comparison in this way cannot verify the efficiency of our method, for standard dataset has not been set up on this problem until now. In fact, it is difficult to compare different works directly since the kinds of genres, the source of samples and the length of clips.

As mentioned in Section 1, the decision tree method is the most used one in genre classification. The motivation of this paper is to perform classification using the reduced feature set based on rough set. To testify the efficiency of feature selection and

Table 2. Classification result based on C4.5 with original feature set

	AM	FB	BB	DOC	CT	MTV	Recall
AM	27	0	0	0	1	2	0.90
FB	0	28	0	2	0	0	0.93
BB	0	0	29	0	1	0	0.97
DOC	0	0	4	25	1	0	0.83
CT	1	0	8	3	16	2	0.53
MTV	4	0	0	4	2	20	0.67
Precision	0.84	1.00	0.70	0.74	0.76	0.83	0.81

classification, we use the C4.5 to do the same work on the same dataset but with the original feature set. The classification result of C4.5 is shown in Table 2.

Compared with the performance generated by C4.5 with original feature set, the proposed approach gets comparative (even slightly better) results. However, the less features that have to be checked, the less time the algorithm consumes, which is the advantage of our methods.

4 Conclusions

Different with existing works, we address the problem of feature selection in video genre classification and propose a rough set approach to accomplish the task. Heuristic search method using rough set theory provides an efficient feature set. Classification using the *rule set* derived from the *reduct* gives good results. The experiments verify its efficiency.

References

1. Fischer, S., Lienhart, R., Effelsberg, W.: Automatic Recognition of Film Genres. In: Proc. of the 3rd ACM Int. Conf. on Multimedia, San Francisco, US, (1995) 295-304
2. Truong, B.T., Dorai, C.: Automatic genre identification for content-based video categorization. In: Proc. of 15th ICPR, (2000) (4) 230-233
3. Liu, Z., Huang, J.C., Wang, Y., Chen, T.: Audio feature extraction and analysis for scene classification. In: Proc. of IEEE Signal Processing Society Workshop on Multimedia Signal Processing, Princeton, US, (1997) 343-348
4. Liu, Z., Huang, J.C., Wang, Y.: Classification of TV programs based on audio information using hidden Markov Model. In: Proc. of IEEE Workshop Multimedia Signal Processing, Los Angeles, US, (1998) 27-32
5. Xu, L.Q., Li, Y.: Video classification using spatial-temporal features and PCA. In: Proc. of the ICME, Baltimore, US, (2003) (3): 485-488
6. Ma, Y.F., Zhang, H.J.: Motion pattern based video classification using support vector machines. EURASIP JASP, 2(2003): 199-208
7. Jin, S.H., Bae, T.M., Choo J.H., Ro, Y.M.: Video genre classification using multimodal features. In: Yeung, M.M.(eds): Storage and Retrieval Methods and Applications for Multimedia 2004. SPIE, Vol. 5307, (2003) 307-318

8. Rasheed, Z., Shah, M.: Movie genre classification by exploiting audio-visual features of previews. In: Proc. of 16th ICPR, Orlando, US, 2(2002):1086-1089
9. Pawlak, Z.: Rough sets: present state and the future. *Foundations of Computing and Decision Sciences*, 11 (1993): 157-166
10. Zhang, M., Yao, J.T.: A rough sets based approach to feature selection. In: Proc. of the 23rd Inter. Conf. of AFIPS, Banff, CA, (2004): 434-439
11. Maaate: <http://www.cmis.csiro.au/dmis/Maaate/>
12. John, G.H., Kohavi, R., Pfleger, K.: Irrelevant features and the subset selection problem. In: Proc. of 11th ICML, San Francisco, CA, (1994):121-129
13. Hu, X.H.: Knowledge discovery in databases: an attribute-oriented rough set approach. Ph.D Dissertation, University of Regina, Canada, 1995
14. Zhong, N., Dong, J.Z., Ohsyga, S.: Using Rough Sets with Heuristics for feature Selection. *Journal of Intelligent Information Systems*, 16(2001):199-214
15. Rosetta: <http://www.idi.ntnu.no/~aleks/rosetta/rosetta.html>

Author Index

- Aboutajdine, Driss 767
Aghajan, Hamid 299
Ahn, Byung-ha 867
Alberich-Carramiñana, Maria 944
Albouy, Benjamin 1111
Alenyà, Guillem 944
Amonou, Isabelle 375
Andrade-Cetto, Juan 944
Antón-Canalis, Luis 501
Antonini, Gianluca 710
Archetti, Francesco 263
Armingol, Jose M. 1151
Astola, Jaakko 46
- Bai, Li 786
Bajramovic, Ferid 1186
Barbarien, Joeri 364
Barcelos, Celia A.Z. 746
Bariamis, Dimitris 67
Barnich, Olivier 734
Batista, Marcos A. 746
Benuskova, Lubica 1133
Béréziat, Dominique 185
Bernard, Thierry M. 220
Bertolino, Pascal 384
Bian, Zhengzhong 887
Blanc-Talon, Jacques 127
Boire, Jean-Yves 909
Bourennane, Salah 127, 1198
Bunyak, Filiz 920
Burrus, Nicolas 220
Butko, Nicholas 1186
- Canchola, Sandra 956
Candela, Vicente F. 24
Carlin, Mats 855
Chang, Ju Yong 598
Charron, Cyril 1099
Chehdi, Kacem 46
Chen, Zhiqing 776
Cheng, Li-Tien 1014
Cheng, Wengang 1210
Cho, A-Young 933
Cho, Ik-Hwan 344, 933
- Cho, Sung-Bae 1080, 1143
Choi, Tae-Young 485
Chung, Tsiung-Iou 654
Clerckx, Tom 418
Collado, Juan M. 1151
Conci, Aura 208
Cornelis, Jan 364, 418, 722
Crnojević, Vladimir S. 103
- Dai, Qionghai 406
Daire, Jean-Luc 897
Danyali, Habibollah 877
Daoudi, Mohamed 767
de A. Vieira, Evelyn 208
De Bock, Johan 232
De Cock, Jan 533
de Groot, Bastiaan 1122
de la Escalera, Arturo 1151
de Melo, Rafael H.C. 208
De Neve, Wesley 442
De Schrijver, Davy 442, 533
De Smet, Patrick 232
de With, Peter H.N. 474, 522
De Witte, Valérie 78
Denzler, Joachim 1186
Dhondt, Yves 442
Di, Huijun 610
Dizdaroglu, Bekir 55
Dong, Haitao 493
Dornaika, Fadi 675
Dosil, Raquel 332
Dubuisson, Séverine 185
- Egiazarian, Karen 46
El Abed, Abir 185
El Fkihi, Sanaa 767
Eom, Seongeun 867
- Falcão, Alexandre X. 138
Fdez-Vidal, Xosé Ramón 332
Fisamen, Britta 855
- Gangal, Ali 55
Gao, Li 323

- García, Antón 332
 Gaucel, Jean-Michel 1198
 Glab, Grzegorz 821
 Gómez-Moreno, Hilario 1163
 Gonzalez-Mora, Jose 1002
 Groen, Frans C.A. 1122
 Groeneweg, Nikolaj J.C. 1122
 Guil, Nicolas 1002
 Guillaume, Mireille 1198

 Halma, Arvid H.R. 1122
 Han, Dongyeob 1026
 Han, Jungong 474
 He, Yuan 846
 Heikkilä, Janne 35
 Hernández-Tejera, Mario 501
 Hilario, Cristina 1151
 Hou, Young-Chang 666
 Hsu, Ching-Sheng 666
 Hu, Dongcheng 846
 Hung, Wen-Liang 654
 Huysmans, Bruno 12
 Hwang, Keum-Sung 1143
 Hyacinthe, Jean-Noël 897

 Iakovidis, Dimitris K. 67, 197
 Im, Seung-Bin 1080
 Intrator, Nathan 554

 Jacob, Jean-Pascal 897
 Jeong, Dong-Seok 344, 933
 Jeong, Jechang 253, 396, 431, 454, 466
 Jeong, Taeuk 545
 Jiang, Gangyi 485, 493
 Jiang, Jianmin 323
 Jiang, Xiaoyue 687
 Jiménez, Hugo 956
 Jin, Zhipeng 485
 Jodogne, Sébastien 734
 Jung, Ho Yub 588

 Karkanis, Stavros A. 197
 Kasabov, Nikola 1133
 Kassim, Ashraf A. 242
 Kerre, Etienne E. 12, 78, 114
 Kim, Chul Keun 1
 Kim, Donghyung 253, 396, 431, 454, 466
 Kim, Hae-Kwang 933
 Kim, Jongho 253, 396, 454, 466
 Kim, Joong Kyu 699
 Kim, Kio 554
 Kim, Sang-Woon 1174
 Kim, Seungjong 431
 Kim, Seungjun 867
 Kim, Taekyung 757
 Kim, Yongil 1026
 Kirkhus, Trine 855
 Kittler, Josef 1037
 Koenig, Emilie 1111
 Kwolek, Bogdan 287

 Labbani-Igbida, Ouiddad 1099
 Lafuente-Arroyo, Sergio 1163
 Lambert, Peter 442, 533
 Landa, Yanina 1014
 Larrey-Ruiz, Jorge 564
 Lau, Frances 299
 Ledda, Alessandro 78
 Lee, Chulhee 545
 Lee, Dongeun 757
 Lee, Inho 275
 Lee, Jaebin 1026
 Lee, Jung-Ho 344
 Lee, Kyoung Mu 588, 598, 978, 990
 Lee, Sang Uk 588, 598, 978, 990
 Lee, Seongwon 757
 Lee, Woong-Ho 344
 Lee, Yung-Ki 311
 Lee, Yung-Lyul 311
 Lemaire, Jean-Jacques 909
 Li, Qian 620
 Liao, Zhihong 406
 Lim, Hong Yin 242
 Lin, Huei-Yung 1047
 Liu, Chang'an 1210
 Liu, Xin U 833
 Liu, Yebin 406
 Liu, Yuncai 576
 Lopes, Ivan O. 746
 López-Sastre, Roberto Javier 1163
 Lu, Hanqing 776
 Lucas, Yves 1111
 Lukin, Vladimir 46
 Luo, Yupin 846
 Luong, Hiệp Q. 78

- Ma, Lihong 776
 Manfredotti, Cristina E. 263
 Maroulis, Dimitris E. 67, 197
 Martínez, Elisa 944
 Martínez, Pedro 956
 Mattern, Frank 1186
 Mertins, Alfred 877
 Messina, Vincenzina 263
 Michel, Bierlaire 710
 Mikulastik, Patrick A. 1059
 Miranda, Paulo A.V. 138
 Mirmehdi, Majid 173
 Moon, Jaekyoung 275
 Moon, Young Shik 799
 Morales-Sánchez, Juan 564
 Morel, Jean-Michel 897
 Mouaddib, El Mustapha 1099
 Munteanu, Adrian 364, 418, 722
- Nath, Sumit K. 920
 Neretti, Nicola 554
 Nicolas, Marina 384
 Nixon, Mark S 833
 Notebaert, Stijn 533
- Oh, Weon-Geun 933
 Ojansivu, Ville 35
 Ostermann, Joern 354
 Oto, Emre 299
 Ouchchane, Lemlih 909
- Paik, Joonki 757
 Palaniappan, Kannappan 920
 Pardo, Xosé Manuel 332
 Park, Bo Gun 978, 990
 Park, Chang-Joon 275
 Park, Gwang Hoon 1
 Park, Hyun 799
 Park, HyunWook 311
 Park, Soon-Yong 275
 Park, Young Kyung 699
 Pesquet-Popescu, Béatrice 375
 Petrović, Nemanja I. 103
 Philips, Wilfried 12, 78, 114, 232
 Pires, Rui 232
 Pižurica, Aleksandra 12
 Pless, Reynaldo C. 956
 Ponomarenko, Nikolay 46
- Quiroga, Bernardo R. 1122
- Rao, Naveed Iqbal 610
 Ravyse, Ilse 810
 Renard, Nadine 127
 Ri, Song-Hak 354
 Richard, Tsai 1014
 Robert, Antoine 375
 Rocha, Anderson 138
 Romero, Pantaleón D. 24
 Roussel, Jérôme 384
 Ruedin, Ana 91
 Ryan, Øyvind 150
- Sahli, Hichem 810
 Salas, Joaquín 956
 Sánchez-Cruz, Hermilo 161
 Sánchez-Nielsen, Elena 501
 Sappa, Angel D. 675
 Savelonas, Michalis A. 197
 Schelkens, Peter 364, 418, 722
 Schulte, Stefan 12
 Schumann-Olsen, Henrik 855
 Seo, Hae Jong 699
 Shao, Feng 493
 Shevchenko, Mikhail 1037
 Shi, Fanhuai 576
 Shin, Vladimir 867
 Shioyama, Tadayoshi 966
 Siegmann, Philip 1163
 Silva, Anselmo M. 746
 Sivertsen, Ronald 855
 Smereka, Marcin 821
 Song, Yi 786
 Sorci, Matteo 710
 Sorrenti, Domenico Giorgio 263
 Stegmann, Ivo 1059
 Stoufs, Maryse R. 722
 Suh, Doug Young 1
- Thielemann, Jens T 855
 Thiran, Jean-Philippe 710
 Torras, Carme 944
 Treuillet, Sylvie 1111
 Tromp, Maarten 1122
 Tu, Shu-Fen 666
- Urfahoglu, Onay 1059
- Vachier, Corinne 897
 Vallée, Jean-Paul 897
 Van de Walle, Rik 442, 533

- Van der Weken, Dietrich 114
Van Deursen, Davy 442
Van Droogenbroeck, Marc 734
Vansteenkiste, Ewout 114
Vázquez-Reina, Amelio 1163
Villéger, Alice C 909
Vozel, Benoit 46
- Wang, Changguo 887
Wang, Guoping 632
Wang, Jianhua 576
Wang, Ming-Liang 1047
Wang, Shi-gang 620
Wang, Xingbo 1210
Wang, Yuzhou 632
Windridge, David 1037
Wu, Fei 513
Wysoski, Simeí Gomes 1133
- Xie, Xianghua 173
Xu, GuangYou 610
Xu, Richard Yi Da 1088
Xu, Wenli 406
- Yang, Jie 644
Yang, Miin-Shen 654
Yang, Ronghua 173
Yang, Shuyuan Y. 323
Yang, You 493
Yang, Yuxiang 887
Yu, Decong 776
Yu, Gang 887
Yu, Kiyun 1026
Yu, Mei 485, 493
Yu, Ping-Hsiu 1047
- Zafarifar, Bahman 522
Zapata, Emilio L. 1002
Zhang, Hongmei 887
Zhang, Jian 513
Zhang, Jing 576
Zhang, Kai 632
Zhao, Jianmin 644
Zhao, Rongchun 687
Zhao, Tuo 687
Zheng, Zhonglong 644
Zhuang, Yueting 513
Zorski, Witold 1071