

Mohamed Kamel
Aurélio Campilho (Eds.)

LNCS 6753

Image Analysis and Recognition

8th International Conference, ICIAR 2011
Burnaby, BC, Canada, June 2011
Proceedings, Part I

1
Part I

 Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

TU Dortmund University, Germany

Madhu Sudan

Microsoft Research, Cambridge, MA, USA

Demetri Terzopoulos

University of California, Los Angeles, CA, USA

Doug Tygar

University of California, Berkeley, CA, USA

Gerhard Weikum

Max Planck Institute for Informatics, Saarbruecken, Germany

Mohamed Kamel Aurélio Campilho (Eds.)

Image Analysis and Recognition

8th International Conference, ICIAR 2011
Burnaby, BC, Canada, June 22-24, 2011
Proceedings, Part I

Volume Editors

Mohamed Kamel
University of Waterloo
Department of Electrical and Computer Engineering
Waterloo, ON, N2L 3G1, Canada
E-mail: mkamel@uwaterloo.ca

Aurélio Campilho
University of Porto
Faculty of Engineering
Institute of Biomedical Engineering
Rua Dr. Roberto Frias
4200-465 Porto, Portugal
E-mail: campilho@fe.up.pt

ISSN 0302-9743
ISBN 978-3-642-21592-6
DOI 10.1007/978-3-642-21593-3
Springer Heidelberg Dordrecht London New York

e-ISSN 1611-3349
e-ISBN 978-3-642-21593-3

Library of Congress Control Number: 2011928908

CR Subject Classification (1998): I.4, I.5, I.2.10, I.2, I.3.5, F.2.2

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

© Springer-Verlag Berlin Heidelberg 2011

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.

The use of general descriptive names, registered names, trademarks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

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

Printed on acid-free paper

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

Preface

This is one of two volumes that contain papers accepted for ICIAR 2011, the International Conference on Image Analysis and Recognition, held at Simon Fraser University, Burnaby, BC, Canada, June 22–24, 2011. This was the eighth edition in the ICIAR series of annual conferences alternating between Europe and North America. The idea of organizing these conferences was to foster collaboration and exchange between researchers and scientists in the broad fields of image analysis and pattern recognition, addressing recent advances in theory, methodology and applications. ICIAR was organized at the same time with AIS 2011, the International Conference on Autonomous and Intelligent Systems. Both conferences were organized by AIMI – Association for Image and Machine Intelligence—a not-for-profit organization registered in Ontario, Canada.

For ICIAR 2011, we received a total of 147 full papers from 37 countries. The review process was carried out by members of the Program Committee and other reviewers; all are experts in various image analysis and pattern recognition areas. Each paper was reviewed by at least two reviewers and checked by the Conference Chairs. A total of 84 papers were finally accepted and appear in the two volumes of these proceedings. The high quality of the papers is attributed first to the authors, and second to the quality of the reviews provided by the experts. We would like to sincerely thank the authors for responding to our call, and to thank the reviewers for their careful evaluation and feedback provided to the authors. It is this collective effort that resulted in the strong conference program and high-quality proceedings.

This year ICIAR had a competition on “Hand Geometric Points Detection,” which attracted the attention of participants.

We were very pleased to be able to include in the conference program keynote talks by well-known experts: Toshio Fukuda, Nagoya University, Japan; William A. Gruver, Simon Fraser University, Canada; Ze-Nian Li, Simon Fraser University, Canada; Andrew Sixsmith, Simon Fraser University, Canada; and Patrick Wang, Northeastern University Boston, USA. We would like to express our sincere gratitude to the keynote speakers for accepting our invitation to share their vision and recent advances in their specialized areas.

Special thanks are also due to Jie Liang, Chair of the local Organizing Committee, and members of the committee for their advice and help. We are thankful for the support and facilities provided by Simon Fraser University. We are also grateful to Springer’s editorial staff for supporting this publication in the LNCS series.

We would like to thank Khaled Hammouda, the webmaster of the conference, for maintaining the Web pages, interacting with the authors and preparing the proceedings.

Finally, we were very pleased to welcome all the participants to ICIAR 2011. For those who did not attend, we hope this publication provides a good view of the research presented at the conference, and we look forward to meeting you at the next ICIAR conference.

June 2011

Mohamed Kamel
Aurélio Campilho

ICIAR 2011 – International Conference on Image Analysis and Recognition

General Chair

Mohamed Kamel
University of Waterloo, Canada
mkamel@uwaterloo.ca

General Co-chair

Aurélio Campilho
University of Porto, Portugal
campilho@fe.up.pt

Local Organizing Committee

Jie Liang (Chair)
Simon Fraser University
Canada

Carlo Menon
Simon Fraser University
Canada

Faisal Beg
Simon Fraser University
Canada

Jian Pei
Simon Fraser University
Canada

Ivan Bajic
Simon Fraser University
Canada

Conference Secretary

Cathie Lowell
Toronto, Ontario, Canada
c.lowell@ieee.org

Webmaster

Khaled Hammouda
Waterloo, Ontario, Canada
khaledh@gmail.com

Supported by



AIMI – Association for Image and Machine Intelligence



PAMI – Pattern Analysis and Machine Intelligence Group
University of Waterloo
Canada



Universidade do Porto

FEUP Faculdade de Engenharia

Department of Electrical and Computer Engineering
Faculty of Engineering
University of Porto
Portugal



INEB – Instituto de Engenharia Biomédica
Portugal

Advisory Committee

M. Ahmadi	University of Windsor, Canada
P. Bhattacharya	Concordia University, Canada
T.D. Bui	Concordia University, Canada
M. Cheriet	University of Quebec, Canada
E. Dubois	University of Ottawa, Canada
Z. Duric	George Mason University, USA
G. Granlund	Linköping University, Sweden
L. Guan	Ryerson University, Canada
M. Haindl	Institute of Information Theory and Automation, Czech Republic
E. Hancock	The University of York, UK
J. Kovacevic	Carnegie Mellon University, USA
M. Kunt	Swiss Federal Institute of Technology (EPFL), Switzerland
J. Padilha	University of Porto, Portugal
K.N. Plataniotis	University of Toronto, Canada
A. Sanfeliu	Technical University of Catalonia, Spain
M. Shah	University of Central Florida, USA
M. Sid-Ahmed	University of Windsor, Canada
C.Y. Suen	Concordia University, Canada
A.N. Venetsanopoulos	University of Toronto, Canada
M. Viergever	University of Utrecht, The Netherlands

B. Vijayakumar	Carnegie Mellon University, USA
J. Villanueva	Autonomous University of Barcelona, Spain
R. Ward	University of British Columbia, Canada
D. Zhang	The Hong Kong Polytechnic University, Hong Kong

Program Committee

A. Abate	University of Salerno, Italy
P. Aguiar	Institute for Systems and Robotics, Portugal
M. Ahmed	Wilfrid Laurier University, Canada
J. Alirezaie	Ryerson University, Canada
H. Araújo	University of Coimbra, Portugal
N. Arica	Turkish Naval Academy, Turkey
I. Bajic	Simon Fraser University, Canada
J. Barbosa	University of Porto, Portugal
J. Barron	University of Western Ontario, Canada
J. Batista	University of Coimbra, Portugal
C. Bauckhage	York University, Canada
G. Bilodeau	École Polytechnique de Montréal, Canada
J. Bioucas	Technical University of Lisbon, Portugal
B. Boufama	University of Windsor, Canada
T.D. Bui	Concordia University, Canada
X. Cao	Beihang University, China
J. Cardoso	University of Porto, Portugal
E. Cernadas	University of Vigo, Spain
M. Cheriet	University of Quebec, Canada
M. Coimbra	University of Porto, Portugal
M. Correia	University of Porto, Portugal
L. Corte-Real	University of Porto, Portugal
J. Costeira	Technical University of Lisbon, Portugal
A. Dawoud	University of South Alabama, USA
M. De Gregorio	Istituto di Cibernetica “E. Caianiello” - CNR, Italy
J. Dias	University of Coimbra, Portugal
Z. Duric	George Mason University, USA
N. El Gayar	Nile University, Egypt
M. El-Sakka	University of Western Ontario, Canada
D. ElShafie	McGill University, Canada
M. Figueiredo	Technical University of Lisbon, Portugal
G. Freeman	University of Waterloo, Canada
L. Guan	Ryerson University, Canada
F. Guibault	École Polytechnique de Montréal, Canada
M. Haindl	Institute of Information Theory and Automation, Czech Republic
E. Hancock	University of York, UK
C. Hong	Hong Kong Polytechnic, Hong Kong
K. Huang	Chinese Academy of Sciences, China

J. Jiang	University of Bradford, UK
J. Jorge	INESC-ID, Portugal
G. Khan	Ryerson University, Canada
M. Khan	Saudi Arabia
Y. Kita	National Institute AIST, Japan
A. Kong	Nanyang Technological University, Singapore
J. Laaksonen	Aalto University, Finland
Q. Li	Western Kentucky University, USA
X. Li	University of London, UK
J. Liang	Simon Fraser University, Canada
R. Lins	Universidade Federal de Pernambuco, Brazil
J. Lorenzo-Ginori	Universidad Central “Marta Abreu” de Las Villas, Cuba
R. Lukac	University of Toronto, Canada
A. Mansouri	Université de Bourgogne, France
A. Marçal	University of Porto, Portugal
J. Marques	Technical University of Lisbon, Portugal
M. Melkemi	Univeristé de Haute Alsace, France
A. Mendonça	University of Porto, Portugal
J. Meunier	University of Montreal, Canada
M. Mignotte	University of Montreal, Canada
A. Mohammed	University of Waterloo, Canada
A. Monteiro	University of Porto, Portugal
M. Nappi	University of Salerno, Italy
A. Padilha	University of Porto, Portugal
F. Perales	University of the Balearic Islands, Spain
F. Pereira	Technical University of Lisbon, Portugal
E. Petrakis	Technical University of Crete, Greece
P. Pina	Technical University of Lisbon, Portugal
A. Pinho	University of Aveiro, Portugal
J. Pinto	Technical University of Lisbon, Portugal
P. Quelhas	Biomedical Engineering Institute, Portugal
M. Queluz	Technical University of Lisbon, Portugal
P. Radeva	Autonomous University of Barcelona, Spain
B. Raducanu	Autonomous University of Barcelona, Spain
S. Rahnamayan	University of Ontario Institute of Technology (UOIT), Canada
E. Ribeiro	Florida Institute of Technology, USA
J. Sanches	Technical University of Lisbon, Portugal
J. Sánchez	University of Las Palmas de Gran Canaria, Spain
B. Santos	University of Aveiro, Portugal
A. Sappa	Computer Vision Center, Spain
A. Sayedelahl	University of Waterloo, Canada
G. Schaefer	Nottingham Trent University, UK
P. Scheunders	University of Antwerp, Belgium

J. Sequeira	Ecole Supérieure d'Ingénieurs de Luminy, France
J. Shen	Singapore Management University, Singapore
J. Silva	University of Porto, Portugal
B. Smolka	Silesian University of Technology, Poland
M. Song	Hong Kong Polytechnical University, Hong Kong
J. Sousa	Technical University of Lisbon, Portugal
H. Suesse	Friedrich Schiller University Jena, Germany
S. Sural	Indian Institute of Technology, India
S. Suthaharan	USA
A. Taboada-Crispí	Universidad Central "Marta Abreu" de las Villas, Cuba
D. Tao	NTU, Singapore
M. Vento	University of Salerno, Italy
Y. Voisin	Université de Bourgogne, France
E. Vrscay	University of Waterloo, Canada
Z. Wang	University of Waterloo, Canada
M. Wirth	University of Guelph, Canada
J. Wu	University of Windsor, Canada
F. Yarman-Vural	Middle East Technical University, Turkey
J. Zelek	University of Waterloo, Canada
L. Zhang	The Hong Kong Polytechnic University, Hong Kong
L. Zhang	Wuhan University, China
Q. Zhang	Waseda University, Japan
G. Zheng	University of Bern, Switzerland
H. Zhou	Queen Mary College, UK
D. Ziou	University of Sherbrooke, Canada

Reviewers

A. Abdel-Dayem	Laurentian University, Canada
J. Ferreira	University of Porto, Portugal
D. Frejlichowski	West Pomeranian University of Technology, Poland
M. Gangeh	University of Waterloo, Canada
S. Mahmoud	University of Waterloo, Canada
A. Mohebi	University of Waterloo, Canada
F. Monteiro	IPB, Portugal
Y. Ou	University of Pennsylvania, USA
R. Rocha	Biomedical Engineering Institute, Portugal

Table of Contents – Part I

Image and Video Processing

Enhancing Video Denoising Algorithms by Fusion from Multiple Views	1
<i>Kai Zeng and Zhou Wang</i>	
Single Image Example-Based Super-Resolution Using Cross-Scale Patch Matching and Markov Random Field Modelling	11
<i>Tijana Ružić, Hiệp Q. Luong, Aleksandra Pižurica, and Wilfried Philips</i>	
Background Images Generation Based on the Nelder-Mead Simplex Algorithm Using the Eigenbackground Model	21
<i>Charles-Henri Quivy and Itsuo Kumazawa</i>	
Phase Congruency Based Technique for the Removal of Rain from Video	30
<i>Varun Santhaseelan and Vijayan K. Asari</i>	
A Flexible Framework for Local Phase Coherence Computation	40
<i>Rania Hassen, Zhou Wang, and Magdy Salama</i>	
Edge Detection by Sliding Wedgelets	50
<i>Agnieszka Lisowska</i>	
Adaptive Non-linear Diffusion in Wavelet Domain	58
<i>Ajay K. Mandava and Emma E. Regentova</i>	
Wavelet Domain Blur Invariants for 1D Discrete Signals	69
<i>Iman Makaremi, Karl Leboeuf, and Majid Ahmadi</i>	
A Super Resolution Algorithm to Improve the Hough Transform	80
<i>Chunling Tu, Barend Jacobus van Wyk, Karim Djouani, Yskandar Hamam, and Shengzhi Du</i>	
Fusion of Multi-spectral Image Using Non-separable Additive Wavelets for High Spatial Resolution Enhancement	90
<i>Bin Liu and Weijie Liu</i>	
A Class of Image Metrics Based on the Structural Similarity Quality Index	100
<i>Dominique Brunet, Edward R. Vrscay, and Zhou Wang</i>	

Structural Fidelity vs. Naturalness - Objective Assessment of Tone
Mapped Images 111
Hojatollah Yeganeh and Zhou Wang

Feature Extraction and Pattern Recognition

Learning Sparse Features On-Line for Image Classification 122
Ziming Zhang, Jiawei Huang, and Ze-Nian Li

Classifying Data Considering Pairs of Patients in a Relational Space 132
Siti Mariam Shafie and Maria Petrou

Hierarchical Spatial Matching Kernel for Image Categorization 141
*Tam T. Le, Yousun Kang, Akihiro Sugimoto, Son T. Tran, and
Thuc D. Nguyen*

Computer Vision

Feature Selection for Tracker-Less Human Activity Recognition 152
Plinio Moreno, Pedro Ribeiro, and José Santos-Victor

Classification of Atomic Density Distributions Using Scale Invariant
Blob Localization 161
*Kai Cordes, Oliver Topic, Manuel Scherer, Carsten Klempt,
Bodo Rosenhahn, and Jörn Ostermann*

A Graph-Kernel Method for Re-identification 173
Luc Brun, Donatello Conte, Pasquale Foggia, and Mario Vento

Automatic Recognition of 2D Shapes from a Set of Points 183
*Benoît Presles, Johan Debayle, Yvan Maillot, and
Jean-Charles Pinoli*

Steganalysis of LSB Matching Based on the Statistical Analysis of
Empirical Matrix 193
Hamidreza Dastmalchi and Karim Faez

Infinite Generalized Gaussian Mixture Modeling and Applications 201
Tarek Elguebaly and Nizar Bouguila

Fusion of Elevation Data into Satellite Image Classification Using
Refined Production Rules 211
Bilal Al Momani, Philip Morrow, and Sally McClean

Using Grid Based Feature Localization for Fast Image Matching 221
Daniel Fleck and Zoran Duric

A Hybrid Representation of Imbalanced Points for Two-Layer Matching	232
<i>Qi Li</i>	
Wide-Baseline Correspondence from Locally Affine Invariant Contour Matching	242
<i>Zhaozhong Wang and Lei Wang</i>	
Measuring the Coverage of Interest Point Detectors	253
<i>Shoaib Ehsan, Nadia Kanwal, Adrian F. Clark, and Klaus D. McDonald-Maier</i>	
Non-uniform Mesh Warping for Content-Aware Image Retargeting	262
<i>Huiyun Bao and Xueqing Li</i>	
Moving Edge Segment Matching for the Detection of Moving Object . . .	274
<i>Mahbub Murshed, Adin Ramirez, and Oksam Chae</i>	
Gauss-Laguerre Keypoints Extraction Using Fast Hermite Projection Method	284
<i>Dmitry V. Sorokin, Maxim M. Mizotin, and Andrey S. Krylov</i>	
Re-identification of Visual Targets in Camera Networks: A Comparison of Techniques	294
<i>Dario Figueira and Alexandre Bernardino</i>	
Statistical Significance Based Graph Cut Segmentation for Shrinking Bias	304
<i>Sema Candemir and Yusuf Sinan Akgul</i>	
Real-Time People Detection in Videos Using Geometrical Features and Adaptive Boosting	314
<i>Pablo Julian Pedrocca and Mohand Saïd Allili</i>	
Color, Texture, Motion and Shape	
A Higher-Order Model for Fluid Motion Estimation	325
<i>Wei Liu and Eraldo Ribeiro</i>	
Dictionary Learning in Texture Classification	335
<i>Mehrdad J. Gangeh, Ali Ghodsi, and Mohamed S. Kamel</i>	
Selecting Anchor Points for 2D Skeletonization	344
<i>Luca Serino and Gabriella Sanniti di Baja</i>	
Interactive Segmentation of 3D Images Using a Region Adjacency Graph Representation	354
<i>Ludovic Paulhac, Jean-Yves Ramel, and Tom Renard</i>	

An Algorithm to Detect the Weak-Symmetry of a Simple Polygon 365
Mahmoud Melkemi, Frédéric Cordier, and Nickolas S. Sapidis

Spatially Variant Dimensionality Reduction for the Visualization of
 Multi/Hyperspectral Images 375
*Steven Le Moan, Alamin Mansouri, Yvon Voisin, and
 Jon Y. Hardeberg*

Tracking

Maneuvering Head Motion Tracking by Coarse-to-Fine Particle Filter . . . 385
Yun-Qian Miao, Paul Fieguth, and Mohamed S. Kamel

Multi-camera Relay Tracker Utilizing Color-Based Particle Filtering . . . 395
Xiaochen Dai and Shahram Payandeh

Visual Tracking Using Online Semi-supervised Learning 406
Meng Gao, Huaping Liu, and Fuchun Sun

Solving Multiple-Target Tracking Using Adaptive Filters 416
B. Cancela, M. Ortega, Manuel G. Penedo, and A. Fernández

From Optical Flow to Tracking Objects on Movie Videos 426
Nhat-Tan Nguyen, Alexandra Branzan-Albu, and Denis Laurendeau

Event Detection and Recognition Using Histogram of Oriented
 Gradients and Hidden Markov Models 436
Chun-hao Wang, Yongjin Wang, and Ling Guan

Author Index 447

Table of Contents – Part II

Biomedical Image Analysis

Arabidopsis Thaliana Automatic Cell File Detection and Cell Length Estimation	1
<i>Pedro Quelhas, Jeroen Nieuwland, Walter Dewitte, Ana Maria Mendonça, Jim Murray, and Aurélio Campilho</i>	
A Machine Vision Framework for Automated Localization of Microinjection Sites on Low-Contrast Single Adherent Cells	12
<i>Hadi Esmaeilsabzali, Kelly Sakaki, Nikolai Dechev, Robert D. Burke, and Edward J. Park</i>	
A Texture-Based Probabilistic Approach for Lung Nodule Segmentation	21
<i>Olga Zinoveva, Dmitriy Zinovev, Stephen A. Siena, Daniela S. Raicu, Jacob Furst, and Samuel G. Armato</i>	
Generation of 3D Digital Phantoms of Colon Tissue	31
<i>David Svoboda, Ondřej Homola, and Stanislav Stejskal</i>	
Using the Pupillary Reflex as a Diabetes Occurrence Screening Aid Tool through Neural Networks	40
<i>Vitor Yano, Giselle Ferrari, and Alessandro Zimmer</i>	
Genetic Snake for Medical Ultrasound Image Segmentation	48
<i>Mohammad Talebi and Ahmad Ayatollahi</i>	
3D-Video-fMRI: 3D Motion Tracking in a 3T MRI Environment	59
<i>José Maria Fernandes, Sérgio Tafula, and João Paulo Silva Cunha</i>	
Classification-Based Segmentation of the Region of Interest in Chromatographic Images	68
<i>António V. Sousa, Ana Maria Mendonça, M. Clara Sá-Miranda, and Aurélio Campilho</i>	

Biometrics

A Novel and Efficient Feedback Method for Pupil and Iris Localization	79
<i>Muhammad Talal Ibrahim, Tariq Mehmood, M. Aurangzeb Khan, and Ling Guan</i>	

Fusion of Multiple Candidate Orientations in Fingerprints	89
<i>En Zhu, Edwin Hancock, Jianping Yin, Jianming Zhang, and Huiyao An</i>	
Fingerprint Pattern and Minutiae Fusion in Various Operational Scenarios	101
<i>Azhar Quddus, Ira Konvalinka, Sorin Toda, and Daniel Asraf</i>	
Fingerprint Verification Using Rotation Invariant Feature Codes	111
<i>Muhammad Talal Ibrahim, Yongjin Wang, Ling Guan, and A.N. Venetsanopoulos</i>	
Can Gender Be Predicted from Near-Infrared Face Images?	120
<i>Arun Ross and Cunjian Chen</i>	
Hand Geometry Analysis by Continuous Skeletons	130
<i>Leonid Mestetskiy, Irina Bakina, and Alexey Kurakin</i>	
Kernel Fusion of Audio and Visual Information for Emotion Recognition	140
<i>Yongjin Wang, Rui Zhang, Ling Guan, and A.N. Venetsanopoulos</i>	
Automatic Eye Detection in Human Faces Using Geostatistical Functions and Support Vector Machines	151
<i>João Dallyson S. Almeida, Aristófanés C. Silva, and Anselmo C. Paiva</i>	
Gender Classification Using a Novel Gait Template: Radon Transform of Mean Gait Energy Image	161
<i>Farhad Bagher Oskuie and Karim Faez</i>	
Person Re-identification Using Appearance Classification	170
<i>Kheir-Eddine Aziz, Djamel Merad, and Bernard Fertil</i>	
Face Recognition	
A Method for Robust Multispectral Face Recognition	180
<i>Francesco Nicolo and Natalia A. Schmid</i>	
Robust Face Recognition After Plastic Surgery Using Local Region Analysis	191
<i>Maria De Marsico, Michele Nappi, Daniel Riccio, and Harry Wechsler</i>	
SEMD Based Sparse Gabor Representation for Eyeglasses-Face Recognition	201
<i>Caifang Song, Baocai Yin, and Yanfeng Sun</i>	

Face Recognition on Low Quality Surveillance Images, by Compensating Degradation	212
<i>Shiva Rudrani and Sukhendu Das</i>	
Real-Time 3D Face Recognition with the Integration of Depth and Intensity Images	222
<i>Pengfei Xiong, Lei Huang, and Changping Liu</i>	
Individual Feature–Appearance for Facial Action Recognition	233
<i>Mohamed Dahmane and Jean Meunier</i>	

Image Coding, Compression and Encryption

Lossless Compression of Satellite Image Sets Using Spatial Area Overlap Compensation	243
<i>Vivek Trivedi and Howard Cheng</i>	
Color Image Compression Using Fast VQ with DCT Based Block Indexing Method	253
<i>Loay E. George and Azhar M. Kadim</i>	
Structural Similarity-Based Affine Approximation and Self-similarity of Images Revisited	264
<i>Dominique Brunet, Edward R. Vrscay, and Zhou Wang</i>	
A Fair P2P Scalable Video Streaming Scheme Using Improved Priority Index Assignment and Multi-hierarchical Topology	276
<i>Xiaozheng Huang, Jie Liang, Yan Ding, and Jiangchuan Liu</i>	
A Novel Image Encryption Framework Based on Markov Map and Singular Value Decomposition	286
<i>Gaurav Bhatnagar, Q.M. Jonathan Wu, and Balasubramanian Raman</i>	

Applications

A Self-trainable System for Moving People Counting by Scene Partitioning	297
<i>G. Percannella and M. Vento</i>	
Multiple Classifier System for Urban Area’s Extraction from High Resolution Remote Sensing Imagery	307
<i>Safaa M. Bedawi and Mohamed S. Kamel</i>	
Correction of Atmospheric Turbulence Degraded Sequences Using Grid Smoothing	317
<i>Rishaad Abdoola, Guillaume Noel, Barend van Wyk, and Eric Monacelli</i>	

A New Image-Based Method for Event Detection and Extraction of Noisy Hydrophone Data	328
<i>F. Sattar, P.F. Driessen, and G. Tzanetakis</i>	
Detection of Multiple Preceding Cars in Busy Traffic Using Taillights . . .	338
<i>Rachana A. Gupta and Wesley E. Snyder</i>	
Road Surface Marking Classification Based on a Hierarchical Markov Model	348
<i>Moez Ammar, Sylvie Le Hégarat-Mascle, and Hugues Mounier</i>	
Affine Illumination Compensation on Hyperspectral/Multiangular Remote Sensing Images	360
<i>Pedro Latorre Carmona, Luis Alonso, Filiberto Pla, Jose E. Moreno, and Crystal Schaaf</i>	
Crevasse Detection in Antarctica Using ASTER Images	370
<i>Tao Xu, Wen Yang, Ying Liu, Chunxia Zhou, and Zemin Wang</i>	
Recognition of Trademarks during Sport Television Broadcasts	380
<i>Dariusz Frejlichowski</i>	
An Image Processing Approach to Distance Estimation for Automated Strawberry Harvesting	389
<i>Andrew Busch and Phillip Palk</i>	
A Database for Offline Arabic Handwritten Text Recognition	397
<i>Sabri A. Mahmoud, Irfan Ahmad, Mohammed Alshayeb, and Wasfi G. Al-Khatib</i>	
Author Index	407

Enhancing Video Denoising Algorithms by Fusion from Multiple Views

Kai Zeng and Zhou Wang

Department of Electrical and Computer Engineering, University of Waterloo
Waterloo, ON, N2L 3G1, Canada
kzeng@engmail.uwaterloo.ca, zhouwang@ieee.org

Abstract. Video denoising is highly desirable in many real world applications. It can enhance the perceived quality of video signals, and can also help improve the performance of subsequent processes such as compression, segmentation, and object recognition. In this paper, we propose a method to enhance existing video denoising algorithms by denoising a video signal from multiple views (front-, top-, and side-views). A fusion scheme is then proposed to optimally combine the denoised videos from multiple views into one. We show that such a conceptually simple and easy-to-use strategy, which we call multiple view fusion (MVF), leads to a computationally efficient algorithm that can significantly improve video denoising results upon state-of-the-art algorithms. The effect is especially strong at high noise levels, where the gain over the best video denoising results reported in the literature, can be as high as 2-3 dB in PSNR. Significant visual quality enhancement is also observed and evidenced by improvement in terms of SSIM evaluations.

Keywords: video denoising, image quality enhancement, image fusion, multiple views.

1 Introduction

Digital video or image sequence has become ubiquitous in our everyday lives. It is critically important to maintain the quality of video at an acceptable level in various application environments such as network visual communications. However, video signals are subject to noise contaminations during acquisition and transmission. Effective *video denoising* algorithms that can remove or reduce the noise is often desired. They not only supply video signals that have better perceptual quality, but also help improve the performance of the subsequent processes such as compression, segmentation, resizing, de-interlacing, and object detection, recognition, and tracking [1].

Existing video denoising algorithms may be roughly classified into three categories. In the first category, the video signal is denoised on a frame-by-frame basis, where all that is needed is a 2D still image denoising algorithm applied to each frame of the video sequence independently. Well-known and state-of-the-art still image denoising algorithms include the Matlab Wiener2D function, Bayes

least square estimation based on Gaussian scale mixture model (BLS-GSM) [2], nonlocal means denoising (NLM) [3], K-SVD method [4], Stein’s unbiased risk estimator-linear expansion of threshold algorithm(SURE-LET) [5], and block matching and 3D transform shrinkage method (BM3D) [6]. For the purpose of video denoising, the major advantage of these approaches is memory efficiency, as no storage of previous frames are necessary in order to denoise the current frame. However, since the correlation between neighboring frames is completely ignored, the denoising process does not make use of all available information and thus cannot achieve the best denoising performance.

In natural video signals, there exists strong correlation between adjacent frames. The second category of video denoising approaches exploited such correlation by incorporating both intra- and inter-frame information. It was found that motion estimation and compensation could further enhance inter-frame correlation [7,8,9]. In [7], a motion estimation algorithm was employed for recursive temporal denoising along estimated motion trajectory. Motion compensation processes had also been incorporated into BLS-GSM and SURE-LET methods, leading to the ST-GSM [8] and video SURE-LET algorithms [9]. In [10], it was claimed that finding single motion trajectory may not be the best choice for video denoising. Instead, multiple similar patches in neighboring frames are found that may not reside along a single trajectory. This is followed by transform and shrinkage based denoising procedures. Perhaps one of the most successful video denoising methods in recent years is the extension of BM3D method for video, namely VBM3D [11], which searches similar patches in both intra- and inter-frames and uses 3D bilateral filtering for noise removal after aggregating the similar patches together.

The third category of denoising algorithms treat video sequences as 3D volumes. The algorithms can operate in the space-time domain by adaptive weighted local averaging [12], 3D order-statistic filtering [13], 3D Kalman filtering [14], or 3D Markov model based filtering [15]. They may also be applied in 3D transform domain, where soft/hard thresholding or Bayesian estimation are employed to eliminate noise, followed by an inverse 3D transform that brings the signal back to the space-time domain. The method in [16] is one such example, where 3D dual-tree complex wavelet transform was employed that demonstrates some interesting and desired properties. Recently, several authors investigated 3D-patch based methods and achieved highly competitive denoising performance [17,18].

Ideally, to make the best use of all available information, the best video denoising algorithms would need to operate in 3D (Category 3). However, when there exists significant motion in the video, direct space-time 3D filtering or 3D transform based approaches are difficult to effectively cover all motion-related video content within local region. Meanwhile, 3D-patch based methods are expensive in finding similar 3D-patches in the 3D volume. By contrast, 2D denoising algorithms that use intra- and/or inter-frame information (Categories 1 and 2) can be made much more efficient, but their performance is restricted by not fully making use of the neighboring pixels in all three dimensions simultaneously.

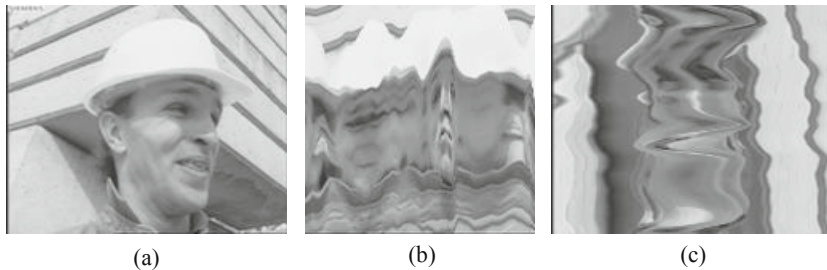


Fig. 1. A video signal observed from (a) front view; (b) side view; and (c) top view

In this paper, we propose a simple strategy, called multiple view fusion (MVF), that provides a useful compromise between 2D (Categories 1 and 2) and 3D (Category 3) approaches. In particular, we denoise the same video volume data with 2D approaches but from three different views, i.e., front view, top view, and side view. An optimal fusion scheme is then employed to combine the three denoised versions of the video. By doing so, the advantage of 2D denoising methods is utilized. Meanwhile, each pixel is denoised by its neighboring pixels from all three dimensions. We show that this simple strategy leads to significant gain of video denoising performance over different base denoising algorithms, especially at high noise levels.

2 Proposed Method

A video signal can be expressed as a 3D function $f(u, v, t)$, where u and v are the horizontal and vertical spatial indices and t is the time index, respectively. A video is typically played along the time axis. At any time instance $t = t_0$, the video is displayed as a 2D front-view image $g_{FV}^{(t_0)}(u, v) = f(u, v, t_0)$ and the image changes over time t . If we think of a video signal as 3D volume data, then it can also be viewed from the side or the top. This gives two other ways to play the same video – a sequence of 2D top-view images $g_{TV}^{(u_0)}(v, t) = f(u_0, v, t)$ for different values of u_0 and a sequence of 2D side-view images $g_{SV}^{(v_0)}(u, t) = f(u, v_0, t)$ for different values of v_0 . An example is given in Fig. 1, where the rarely observed side- and top-view images demonstrate some interesting regularized spatiotemporal structures.

Let x be an original noise-free video signal, which is contaminated by additive noise n , resulting in a noisy signal

$$y = x + n. \quad (1)$$

A video denoising operator D takes the noisy observation y and maps it to an estimator of x :

$$\hat{x} = D(y), \quad (2)$$

such that the difference between x and \hat{x} is as small as possible. How to quantify the difference between x and \hat{x} is another subject of study. The most typically

used ones are the mean squared error (MSE) and equivalently the peak-signal-to-noise ratio (PSNR). However, recent studies showed that the structural similarity index (SSIM) [19] may be a better measure in predicting perceived image distortion.

The proposed MVF method relies on a base video denoising algorithm (which could be as simple as frame-by-frame Winer2D, or as complicated as VBM3D [11]). The base denoiser is applied to the same noisy signal y multiple times but from different views, which gives multiple versions of denoised signal

$$\begin{aligned} z_1 &= D_1(y), \\ z_2 &= D_2(y), \\ &\dots, \\ z_N &= D_N(y). \end{aligned} \quad (3)$$

In this paper $N = 3$, as we have three different views, but in principle the general approach also applies to the cases of less or more views, or multiple denoising algorithms. Let $\mathbf{z} = [z_1, z_2, \dots, z_N]^T$ be a vector that contains all denoised results, then the final denoised signal \hat{x} is given by applying a fusion operator F to \mathbf{z} :

$$\hat{x} = D(y) = F(\mathbf{z}) = F(D_1(y), D_2(y), \dots, D_N(y)). \quad (4)$$

In the case that the base denoisers are predetermined, all the remaining task is to define the fusion rule F , which would be desired to achieve certain optimality. Here we employ a weighted average fusion method given by

$$\hat{x} = \mathbf{w}^T(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}}) + \mu_x, \quad (5)$$

where $\mu_x = \mathbb{E}(x)$ (we use \mathbb{E} to denote the expectation operator), $\boldsymbol{\mu}_{\mathbf{z}}$ is a column vector of expected values $[\mathbb{E}(z_1), \mathbb{E}(z_2), \dots, \mathbb{E}(z_N)]^T$, and \mathbf{w} is a column vector $[w_1, w_2, \dots, w_N]^T$ that defines the weight assigned to each denoised signal. To find the optimal weights \mathbf{w} in the least-square sense, we define the following error energy function

$$E = \mathbb{E}[(x - \hat{x})^2] + \lambda \|\mathbf{w} - \frac{1}{N}\mathbf{1}\|^2, \quad (6)$$

where $\mathbf{1}$ is a length- N column vector with all entries equaling 1. The second term is to regularize the weighting vector towards all equal weights, and the parameter λ is used to control the strength of regularization. Taking the derivative of E with respect to \mathbf{w} and setting it zero, we obtain

$$(\mathbf{C}_{\mathbf{z}} + \lambda\mathbf{I})\mathbf{w} = \mathbf{b} + \frac{\lambda}{N}\mathbf{1}, \quad (7)$$

where \mathbf{I} denotes the $N \times N$ identity matrix, $\mathbf{C}_{\mathbf{z}}$ is the covariance matrix

$$\mathbf{C}_{\mathbf{z}} = \mathbb{E}[(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}})(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}})^T], \quad (8)$$

and \mathbf{b} is a column vector given by

$$\mathbf{b} = \mathbb{E}[(x - \mu_x)(\mathbf{z} - \boldsymbol{\mu}_{\mathbf{z}})]. \quad (9)$$

We can then solve for optimal \mathbf{w} , which gives

$$\mathbf{w}_{opt} = (\mathbf{C}_z + \lambda \mathbf{I})^{-1} \left(\mathbf{b} + \frac{\lambda}{N} \mathbf{1} \right). \quad (10)$$

Here the $\lambda \mathbf{I}$ term plays an important role in stabilizing the solution, especially when \mathbf{C}_z is close to singular. The computation of \mathbf{b} requires the original signal x , which is not available. But by assuming n to be zero-mean and independent of \mathbf{z} , we have

$$\mathbf{b} = \mathbb{E}[(y - n - \mu_x)(\mathbf{z} - \boldsymbol{\mu}_z)] = \mathbb{E}[(y - \mu_y)(\mathbf{z} - \boldsymbol{\mu}_z)]. \quad (11)$$

When applying the above approach to real signals, the expectation operators would need to be replaced by sample means. In our implementation, we apply the weight calculation to individual non-overlapping $8 \times 8 \times 8$ blocks, resulting in block-wise space-time adaptive weights in the 3D volume. Eq. (5) is then applied to each block to obtain the final denoised signal.

3 Experimental Result

We use publicly available video sequences to test the proposed algorithm, which include ‘‘Akiyo’’, ‘‘Carphone’’, ‘‘Forman’’, ‘‘Miss America’’, ‘‘News’’, and ‘‘Salesman’’. The size of all sequences is $144 \times 176 \times 144$. Independent white Gaussian noise was added to the original video sequences, where the noise standard deviation, σ , covers a wide range between 10 and 100. All sequences are in YCrCb 4:2:0 format, but only the denoising results of the luma channel was reported here to validate the algorithm. Two objective criteria, namely PSNR and SSIM [19], were employed to evaluate the quality of denoised video quantitatively. PSNR is the most widely used method in the literature, but SSIM has been recognized as a much better measure to predict subjective quality measurement.

Many state-of-the-art denoising algorithms are publicly available that facilitate direct comparisons. Due to space limit, here we report our comparison results for 5 noise levels (σ equals 10, 15, 20, 50, and 100, respectively) using three base denoising methods with and without using our MVF approach. The base algorithms are Matlab Wiener2D, BLS-GSM [2] and VBM3D [11]. We have also applied our MVF approach to a list of other highly competitive algorithms, including NLM [10], K-SVD [4], and SURE-LET [9]. Similar results were obtained but are not reported here.

Table 1 shows the comparison results using PSNR and SSIM measures, which were computed frame-by-frame and then averaged over all frames. It can be seen that the proposed MVF approach consistently leads to performance gain over all base denoising algorithms, for all test video sequences, and at all noise levels. The gain is especially significant at high noise levels, where the improvement can be as high as 2-3 dB in terms of PSNR over state-of-the-art algorithms such as VBM3D, which is among the best algorithms ever reported in the literature. To

Table 1. PSNR and SSIM comparisons for three video denoising algorithms with and without MVF

Video Sequence	<i>Akiyo</i>					<i>Carphone</i>				
Noise std (σ)	10	15	20	50	100	10	15	20	50	100
PSNR Results (dB)										
Wiener-2D	33.22	30.38	28.33	21.58	15.94	32.66	29.84	27.86	21.35	15.86
with MVF	34.69	31.91	29.89	23.15	17.52	33.90	31.20	29.29	22.87	17.42
BLG-GSM	36.12	33.73	32.09	27.32	24.36	35.34	33.00	31.40	26.47	23.15
with MVF	39.95	37.58	35.88	30.78	27.43	37.01	34.92	33.50	29.02	25.81
VBM3D	42.01	39.76	37.91	30.79	24.39	38.50	36.64	35.35	29.82	23.30
with MVF	42.33	40.08	38.36	32.64	26.93	38.50	36.71	35.46	30.97	25.76
SSIM Results										
Wiener-2D	0.876	0.788	0.700	0.364	0.164	0.885	0.803	0.722	0.408	0.205
with MVF	0.906	0.833	0.757	0.432	0.213	0.909	0.840	0.771	0.472	0.255
BLG-GSM	0.952	0.924	0.898	0.765	0.636	0.951	0.927	0.902	0.773	0.627
with MVF	0.977	0.964	0.949	0.866	0.749	0.964	0.947	0.930	0.839	0.718
VBM3D	0.983	0.976	0.965	0.874	0.616	0.972	0.961	0.951	0.874	0.628
with MVF	0.986	0.978	0.967	0.903	0.684	0.972	0.961	0.952	0.892	0.691
Video Sequence	<i>Foreman</i>					<i>Miss America</i>				
PSNR Results (dB)										
Wiener-2D	32.22	29.49	27.55	21.17	15.77	34.36	31.35	29.17	21.91	16.07
with MVF	33.11	30.53	28.70	22.59	17.30	35.74	32.80	30.67	23.47	17.65
BLG-GSM	34.22	31.92	30.32	25.44	22.21	38.69	36.54	35.09	30.61	27.52
with MVF	35.83	33.65	32.12	27.36	24.05	41.03	38.99	37.59	33.16	30.02
VBM3D	37.37	35.50	34.12	28.47	22.46	41.93	40.19	38.81	33.55	26.57
with MVF	37.68	35.80	34.44	29.28	24.14	42.34	40.57	39.24	34.69	28.93
SSIM Results										
Wiener-2D	0.887	0.812	0.738	0.432	0.220	0.848	0.737	0.633	0.275	0.107
with MVF	0.906	0.843	0.778	0.488	0.267	0.879	0.785	0.692	0.331	0.138
BLG-GSM	0.938	0.910	0.884	0.746	0.591	0.958	0.939	0.922	0.841	0.751
with MVF	0.952	0.930	0.908	0.792	0.646	0.972	0.960	0.948	0.884	0.791
VBM3D	0.961	0.947	0.933	0.844	0.601	0.976	0.968	0.959	0.901	0.669
with MVF	0.962	0.948	0.934	0.857	0.643	0.978	0.970	0.962	0.915	0.685
Video Sequence	<i>News</i>					<i>Salesman</i>				
PSNR Results (dB)										
Wiener-2D	31.95	29.11	27.14	20.83	15.65	31.48	28.97	27.23	21.28	15.90
with MVF	33.34	30.58	28.66	22.44	17.26	33.07	30.65	28.94	22.92	17.50
BLG-GSM	34.34	31.86	30.11	24.90	21.42	33.16	30.89	29.37	25.35	23.01
with MVF	37.72	35.30	33.57	28.22	24.58	36.82	34.43	32.82	28.34	25.71
VBM3D	39.76	37.47	35.73	28.50	21.69	38.93	36.49	34.57	27.92	23.18
with MVF	40.04	37.73	36.06	30.18	24.67	39.27	36.84	35.06	29.58	25.52
SSIM Results										
Wiener-2D	0.887	0.807	0.731	0.431	0.231	0.876	0.798	0.724	0.415	0.194
with MVF	0.915	0.851	0.787	0.503	0.292	0.912	0.854	0.796	0.511	0.265
BLG-GSM	0.950	0.923	0.894	0.737	0.564	0.908	0.854	0.804	0.613	0.478
with MVF	0.973	0.958	0.942	0.844	0.712	0.958	0.930	0.902	0.769	0.643
VBM3D	0.981	0.971	0.960	0.860	0.581	0.975	0.956	0.929	0.739	0.488
with MVF	0.982	0.973	0.963	0.895	0.684	0.976	0.958	0.936	0.803	0.618

demonstrate the performance improvement for individual video frames, Fig. 2 depicts PSNR and SSIM comparisons as functions of frame number for “Foreman” sequence. Again, consistent improvement is observed for almost all frames, indicating the robustness of the proposed MVF approach.

Figure 3 provides visual comparisons of the denoising results of one frame extracted from the “Salesman” sequence. For each denoised frame, the SSIM quality map is also given, where brighter pixels indicate higher SSIM values and thus better quality. Visual quality improvement by the proposed MVF approach can be perceived in various locations in the denoised frames, for example, the bookshelf region. Such improvement is also clearly indicated by the SSIM maps.

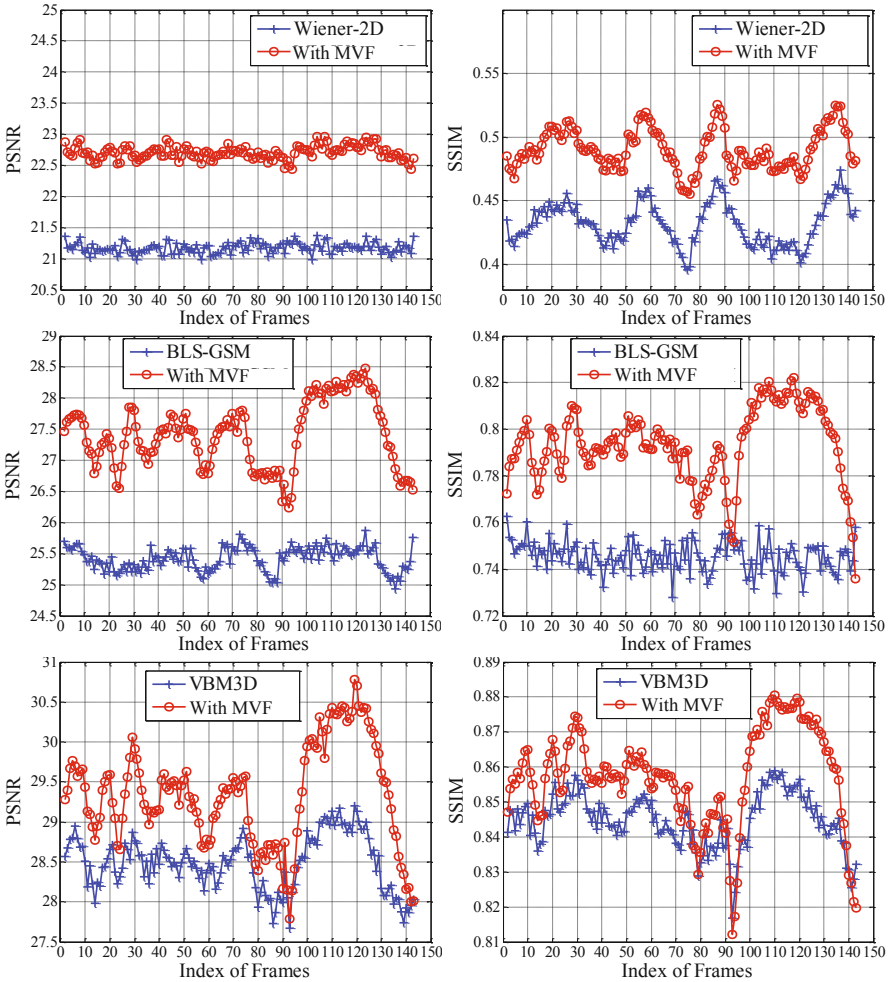


Fig. 2. PSNR and SSIM comparisons as functions of frame number for “Foreman” sequence. Noise level $\sigma = 50$

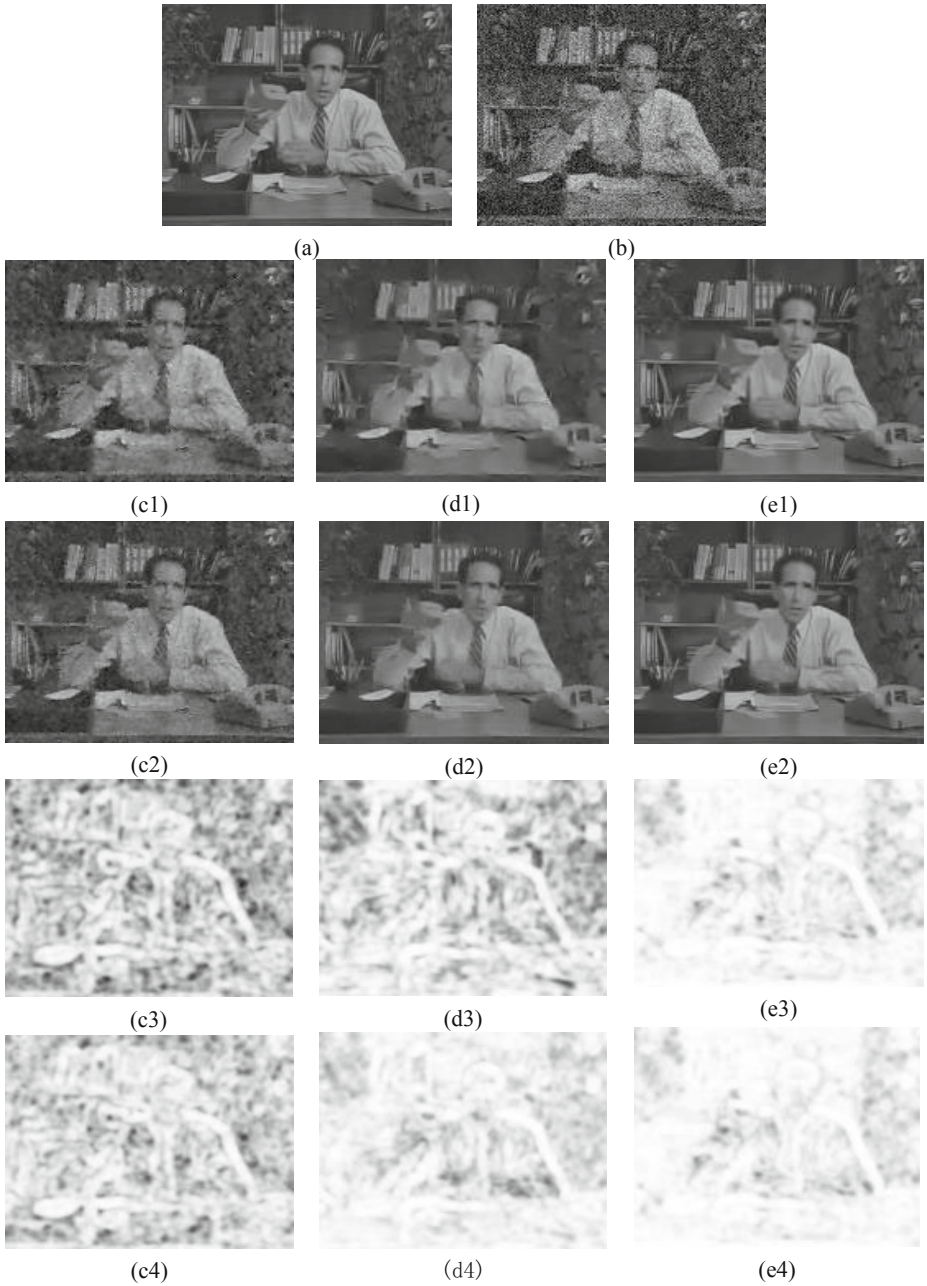


Fig. 3. (a): One frame extracted from original “Salesman” sequence; (b): Corresponding noisy frame with $\sigma = 50$; (c1) to (e1): Wiener2D, BLS-GSM, and VBM3D denoised frames; (c2) to (e2): Wiener2D, BLS-GSM, and VBM3D denoised frames with optimal MVF; (c3) to (e3): SSIM quality maps for (c1) to (e1); (c4) to (e4): SSIM quality maps for (c2) to (e2)

4 Conclusion

We propose an MVF approach that can improve video denoising performance of existing algorithms by fusing the denoising results from multiple views. Our experimental results demonstrate consistent improvement over some of the best video denoising algorithms in the literature. The proposed method is conceptually simple, easy-to-use, and computationally efficient. The complexity of the whole algorithm mainly depends on that of the base denoising method, but not the MVF procedure. In principle, the MVF strategy could be applied to any existing video denoising algorithm, but our major intension here is to apply it to 2D approaches (Categories 1 and 2 described in Section 1). The reason is that the denoising results obtained by applying 2D approaches from different views tend to be complementary to each other. By contrast, 3D approaches (Category 3) such as those using 3D patches have already considered the dependencies between neighboring pixels from all directions, and thus applying them from different views may lead to similar results that would not complement each other to a significant extent.

The video denoising performance may be further improved by adopting better base denoising algorithms or by improving the fusion method. One could also attempt to fuse the denoising results not only from multiple views but also by multiple algorithms. It is also interesting to look into novel algorithms for denoising from side- and top-views, where we have observed special regularities (that are quite different from what has been observed from front-view) that are worth deeper investigations.

Acknowledgment

This research was supported in part by Natural Sciences and Engineering Research Council of Canada in the forms of Discovery, Strategic and CRD Grants, and by an Ontario Early Researcher Award, which are gratefully acknowledged.

References

1. Bovik, A.C.: Handbook of Image and Video Processing (Communications, Networking and Multimedia). Academic Press, Inc., Orlando (2005)
2. Portilla, J., Strela, V., Wainwright, M.J., Simoncelli, E.P.: Image Denoising Using Scale Mixtures of Gaussians in the Wavelet Domain. *IEEE Trans. on Image Processing*. 12, 1338–1351 (2003)
3. Buades, A., Coll, B., Morel, J.M.: Nonlocal Image and Movie Denoising. *Int. J. of Computer Vision* 76, 123–139 (2008)
4. Aharon, M., Elad, M., Bruckstein, A.: K-SVD: An Algorithm for Designing Overcomplete Dictionaries for Sparse Representation. *IEEE Trans. on Signal Processing* 11, 4311–4322 (2006)
5. Blu, T., Luisier, F.: The SURE-LET Approach to Image Denoising. *IEEE Trans. on Image Processing* 16, 2778–2786 (2007)

6. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image Denoising by Sparse 3-D Transform-Domain Collaborative Filtering. *IEEE Trans. on Image Processing*. 16, 2080–2095 (2007)
7. Zlokolica, V., Pizurica, A., Philips, W.: Wavelet-Domain Video Denoising Based on Reliability Measures. *IEEE Trans. on Cir. and Sys. for Video Tech.* 16, 993–1007 (2006)
8. Varghese, G., Wang, Z.: Video Denoising Based on a Spatiotemporal Gaussian Scale Mixture Model. *IEEE Trans. on Cir. and Sys. for Video Tech.* 20, 1032–1040 (2010)
9. Luisier, F., Blu, T., Unser, M.: SURE-LET for Orthonormal Wavelet-Domain Video Denoising. *IEEE Trans. on Cir. and Sys. for Video Tech.* 20, 913–919 (2010)
10. Buades, A., Coll, B., Morel, J.M., Matemàtiques, D.: Denoising Image Sequences does not Require Motion Estimation. In: *Proc. of the IEEE Conf. on Advanced Video and Signal Based Surveillance*, pp. 70–74 (2005)
11. Dabov, K., Foi, A., Egiazarian, K.: Video Denoising by Sparse 3D Transform-Domain Collaborative Filtering. In: *Proc. of the 15th Euro. Signal Proc. Conf., Poland (September 2007)*
12. Ozkan, M.K., Sezan, M.I., Tekalp, A.M.: Adaptive Motion-compensated Filtering of Noisy Image Sequences. *IEEE Trans. on Cir. and Sys. for Video Tech.* 3, 277–290 (1993)
13. Arce, G.R.: Multistage Order Statistic Filters for Image Sequence Processing. *IEEE Trans. on Signal Processing* 39, 1146–1163 (1991)
14. Kim, J., Woods, J.W.: Spatiotemporal Adaptive 3-D Kalman Filter for Video. *IEEE Trans. on Image Processing* 6, 414–424 (1997)
15. Brailean, J.C., Katsaggelos, A.K.: Recursive Displacement Estimation and Restoration of Noisy-blurred Image Sequences. In: *IEEE Int. Conf. on Acoustics, Speech, and Signal Processing*, vol. 5, pp. 273–276 (April 1993)
16. Selesnick, W.I., Li, K.Y.: Video Denoising using 2D and 3D Duality Complex Wavelet Transforms. In: *Proc. SPIE, Wavelets: Applications in Signal and Image Processing X*, vol. 5207, pp. 607–618 (November 2003)
17. Protter, M., Elad, M.: Image Sequence Denoising via Sparse and Redundant Representations. *IEEE Trans. on Image Processing* 18, 27–35 (2009)
18. Li, X., Yunfei, Z.: Patch-based video processing: a variational Bayesian approach. *IEEE Trans. on Cir. and Sys. for Video Tech.* 19, 27–40 (2009)
19. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. on Image Processing* 13, 600–612 (2004)

Single Image Example-Based Super-Resolution Using Cross-Scale Patch Matching and Markov Random Field Modelling

Tijana Ružić, Hiệp Q. Luong, Aleksandra Pižurica, and Wilfried Philips

Ghent University, TELIN-IPI-IBBT,
Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
tijana.ruzic@telin.ugent.be

Abstract. Example-based super-resolution has become increasingly popular over the last few years for its ability to overcome the limitations of classical multi-frame approach. In this paper we present a new example-based method that uses the input low-resolution image itself as a search space for high-resolution patches by exploiting self-similarity across different resolution scales. Found examples are combined in a high-resolution image by the means of Markov Random Field modelling that forces their global agreement. Additionally, we apply back-projection and steering kernel regression as post-processing techniques. In this way, we are able to produce sharp and artefact-free results that are comparable or better than standard interpolation and state-of-the-art super-resolution techniques.

Keywords: Super-resolution, self-similarities, Markov Random Field, kernel regression.

1 Introduction

Super-resolution (SR) plays an important role in image processing applications nowadays due to the huge amount of low resolution video and image material. Low resolution is a consequence of using low-cost imaging sensors for image/video acquisition, such as webcams, cell phones and surveillance cameras. Furthermore, the increasing popularity of HDTV makes the SR methods necessary for resolution enhancement of NTSC and PAL recordings.

The task of SR is to infer a high-resolution image from one or more low resolution images. Among many SR techniques, two approaches can be identified: classical and example-based approach. Classical SR methods attempt to reconstruct a high-resolution (HR) image from a sequence of degraded low-resolution (LR) images taken from the same scene at sub-pixel shifts [1,2]. Each output pixel is related to one or more input pixels by the acquisition or degradation model. If there is an insufficient number of LR images, prior knowledge can be used as an additional source of information. Classical SR in practice, however, is limited to the the magnification factor smaller than two [4]. Example-based

SR is able to overcome this limitation. The goal of the example-based approach [3,4] is to fill in the missing high frequencies by searching for highly similar patches in the external database that also contains high-resolution information. The method actually consists of two steps: a learning and a reconstruction step. The former involves searching for k nearest neighbours in the database for each LR patch of the input image, while the latter combines the corresponding HR patches of those nearest neighbours to form the HR image.

A problem with example-based methods is that they involve storing and searching large databases. Searching the database can be avoided by using it only to learn the interpolation functions [5,6], but still this external database is necessary. Additionally, it is not guaranteed that the database contains the true high-resolution details which may cause the so called “hallucination” effect. Furthermore, this database needs to be large enough to provide good results which makes learning or searching computationally more demanding.

A solution to the previously mentioned problems is to use the LR input image itself as a search space in the learning phase, as implied originally in [3]. Based on this idea, several single image super-resolution techniques have been developed [7,8]. In [8], all examples are obtained by searching for nearest neighbours within the Gaussian pyramid of the input LR image. This example-based part is combined with the classical SR approach to yield an HR image with an arbitrary magnification factor. The use of a single image is justified by the level of patch redundancy within the same scale and across different levels of Gaussian pyramid. Following this reasoning, we have also developed a single-image super-resolution algorithm that, in addition to these non-local similarities within and across scales, uses sparsity constraints to perform image super-resolution [9].

In this paper we propose a novel single image example-based super-resolution algorithm which combines the learning phase of [8] by searching for examples within the Gaussian pyramid of the input image itself and the reconstruction phase of [3], which uses the Markov Random Field (MRF) model to reconstruct the HR image. The main benefit of such learning approach is that no external database is required which results in faster search and absence of “hallucination” effect (when compared with [3]). On the other hand, using MRF in the reconstruction enables us to stay in the example-based domain without combining it with classical SR as in [8]. There are a few advantages to this in comparison with [8]. First of all, we can use only one level of the pyramid as the search space whose sub-sampling factor corresponds to the magnification factor instead of multiple levels with non-integer sub-sampling factor and, thus, again decrease the computation time. Second, we reconstruct only the HR image of the desired resolution rather than employing course-to-fine reconstruction of images at intermediate resolutions. Finally, we avoid sub-pixel registration which often causes inaccurate results.

Another contribution of this paper is that we show that a simpler and faster method can be used for inference in MRF instead of belief propagation used originally in [3]. We use our method from [10] called neighbourhood-consensus

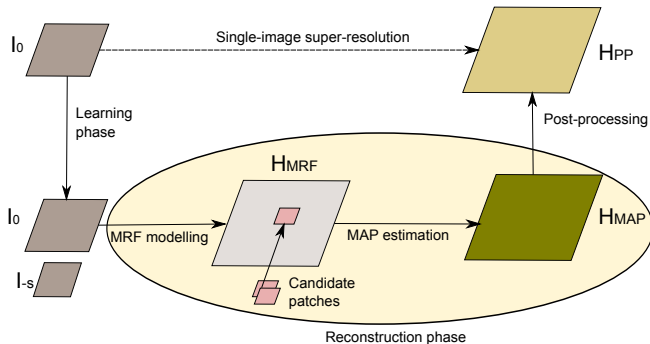


Fig. 1. The proposed single-image example-based super-resolution method

message passing. Our results of the complete algorithm on different test images demonstrate a comparable or better performance than state-of-the-art SR techniques.

This paper is organized as follows. Sec. 2 describes our method for single image example-based SR, where Sec. 2.1 explains in more details the learning phase and Sec. 2.2 the reconstruction phase. Finally, we present and discuss the experimental results in Sec. 3 and we give our conclusion in Sec. 4.

2 Proposed Single Image Example-Based Method

We propose a single-image example-based super-resolution method which uses MRF to model the HR image as a collection of overlapping HR patches whose possible candidates are obtained from the input LR image itself. The algorithm can be divided into three main phases: learning, reconstruction and post-processing (see Fig. 1). In the *learning* phase, we find candidate patches of each unknown HR patch by first searching for k nearest neighbours of its corresponding known LR patch from the input image. This search exploits the patch redundancy across different scales of the Gaussian pyramid. We then extract the HR pairs of the found neighbours (called “parent” patches) from the input image and we use them as candidate patches for corresponding locations in the HR image, because we assume that the LR and HR patches are related in the same way across different scales. What follows is the *reconstruction* phase, which models the HR image as a MRF and performs inference on this model. MRF model has a great advantage over the simpler alternative, i.e. choosing the best match at each location, as we will demonstrate shortly.

Finally, we apply post-processing techniques to eliminate remaining artefacts. We use back-projection [1] to ensure the consistency of the HR result with the input LR image. In case of a small input image and high magnification factor, the search space may become too small for good matches to be found. This will result in visible artefacts so we also use steering kernel regression [11] that produces a smooth and artefact-free image while still preserving edges, ridges

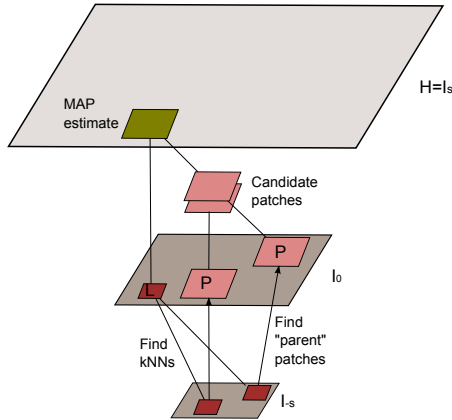


Fig. 2. An illustration of the process of learning candidate patches

and blobs. Post-processing together with MRF modelling allows us to obtain competitive SR result even with only having LR image as the algorithm’s input.

In the remaining of this section we will describe in details the learning and reconstruction phase.

2.1 Learning Candidate Patches

In this section we will explain how to use the single input image to obtain candidate patches. We use the example-based part of the algorithm from [8] in the sense that we search for similar patches within the Gaussian pyramid and use their “parent” HR patches for further reconstruction. However, our approach differs in the reconstruction step which enables us to perform simplified and faster search. Specifically, we search in only one level of the Gaussian pyramid whose sub-sampling factor is equal to the magnification factor for the reasons that will be explained shortly.

We start from the LR input image I_0 which is then blurred and sub-sampled with the integer factor s to yield the lower level of the Gaussian pyramid I_{-s} . The final goal is to reconstruct the image with a resolution that is s times higher than the original resolution. We will denote this HR image with $H = I_s$. The image I_{-s} will serve as a search space for matches of each patch from the image I_0 . In details, the search and matching process has the following course, as illustrated in Fig. 2. For each pixel $p \in I_0$, where $p = (x, y)$ actually represents coordinates of the pixel in the image grid, we take its surrounding patch o_p (denoted by L on Fig. 2) and search for its k nearest neighbours (kNNs) in the image I_{-s} . Those neighbours are the patches that have the lowest sum of squared differences with o_p . Once kNNs y_p^n , $n = 1, \dots, k$, are found, their “parent” patches x_p^n , $n = 1, \dots, k$, (denoted by P on Fig. 2) are extracted from the given image I_0 .

The “parent” patch represents a HR component of the HR-LR pair, where LR component is the LR patch. If the location of the central pixel of the LR

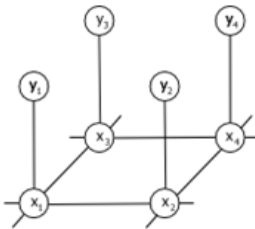


Fig. 3. MRF model: x_p are unknown HR patches and y_p measured LR patches

patch y_p^n (found kNN) is $\tilde{p} \in I_{-s}$, then the location of the central pixel of the “parent” patch x_p^n is $\tilde{sp} \in I_0$ and its size is s times the size of the LR patch. These parent patches can now serve as candidate patches for each location sp in the HR image H which corresponds to the starting location p in the input LR image I_0 . This is the reason for the same value s of the magnification and sub-sampling factor.

2.2 High-Resolution Image Reconstruction

After the algorithm described in the previous subsection, we have k candidate patches x_p^n , $n = 1, \dots, k$, for each location $sp \in H$. These locations correspond to starting locations $p \in I_0$ so we will refer to them with the index p . They are s pixels apart from each other in each direction in the HR grid. The naive approach would be to choose the best match, i.e. the nearest neighbour, at each location. Since the neighbouring patches will normally overlap, we can simply take the average in the overlap region. Although this solution could speed up the search process (because we only search for one nearest neighbour), the resulting image will have visible artefacts (Fig. 4).

Instead of just choosing the HR patch based on its agreement with the available data (the input image), we can take into account the relationship that inevitably exists between neighbouring locations in H in the sense that neighbouring patches should agree in the overlap region. This means that the sum of squared differences in the overlap region is minimal. Furthermore, we would like to observe the image as a whole rather than a collection of local assumptions. In this respect, we can formulate the choice of patches as a global optimization problem over the whole HR image by using the MRF framework [12]. For this purpose, we adopt the concept of [3] with a few major differences. First of all, our candidate patches are obtained from the input image itself, without using an external database. Moreover, they consist of raw pixel values instead of high frequency details so there is no need for preprocessing of the search space. Finally, we use our inference method for optimization which is simpler and faster than loopy belief propagation (LBP) [13] which was originally used.

Specifically, we model H as an undirected graph (Fig. 3) whose hidden nodes, indexed by p , represent the overlapping HR patches in the HR image that can take one of the values from the set $\{x_p\}$. Each hidden node is connected to the observed node (measured data) which is the LR patch o_p around pixel $p \in I_0$.

To completely define MRF model we still have to define compatibility functions between observed and hidden nodes (so called local evidence) and neighbouring hidden nodes. The former determines how much the unknown data agrees with measured data and the latter encodes prior information on the distribution of the unknown image. Local evidence is taken to be the Gaussian function of the matching error, i.e. sum of squared differences, between starting LR patch o_p and found k nearest neighbours y_p^n :

$$\phi_p(y_p^n, o_p) = \exp(-\|y_p^n - o_p\|^2/2\sigma_R^2), \quad (1)$$

Compatibility between neighbouring hidden nodes is the Gaussian function of the matching error in the region of overlap ROV of two neighbouring HR patches:

$$\psi_{p,q}(x_p^n, x_q^m) = \exp(-\|\text{ROV}_{q,p}^n - \text{ROV}_{p,q}^m\|^2/2\sigma_N^2). \quad (2)$$

σ_R and σ_N are the noise covariances which represent the difference between some “ideal” training samples and our image and training samples, respectively. Now, we have to choose one patch from the candidate set at each node that best fits the above constraints over the whole graph. This can be achieved by finding maximum a posteriori (MAP) estimates:

$$\hat{\mathbf{H}} = \hat{\mathbf{x}} = \arg \max_{\mathbf{x}} P(\mathbf{x}|I_0) \quad (3)$$

$$P(\mathbf{x}|I_0) \propto \prod_{p,q} \psi_{p,q}(x_p^n, x_q^m) \prod_p \phi_p(y_p^n, o_p), \quad (4)$$

where ϕ_p is defined in equation 1 and $\psi_{p,q}$ in equation 2. This is generally a difficult problem to be solved exactly, but there is a number of approximate inference algorithms that can yield an approximate solution. We use our inference method called neighbourhood-consensus message passing (NCMP) [10] which is simpler and faster than LBP while the results are qualitatively very similar. Comparison of different approaches for HR image reconstruction is shown in



Fig. 4. Cropped version of zebra image 2x magnification. From left to right: best match result, MRF result with LBP as inference method, MRF result with NCMP as inference method.

Fig. 4. On the left we see the result of the best match approach which has a lot of artefacts due to its greedy nature. Using a MRF model produces much better result even if we use a simple inference method like NCMP (right image).

3 Experimental Results

We tested our method on several images and compared it to the standard interpolation technique, like bi-cubic interpolation, and state-of-the-art SR techniques from [8], which is another single-image SR method, and [14], which uses a parametric learned edge model. In all experiments the LR patch size was 3×3 and HR patch size $3s \times 3s$, while parameters of MRF compatibility functions σ_R and σ_N slightly varied over different images. The number of nearest neighbours was $k = 10$.

In the first experiment, we demonstrate the effectiveness of our technique for sufficiently large search space. Fig. 5 shows the castle image and the results of our super-resolution algorithm with the magnification $s = 2$. It can be seen that the



Fig. 5. Cropped castle image 2x magnification. From left to right and top to bottom: bi-cubic interpolation, MRF result, MRF with back-projection, MRF with back-projection and kernel regression.



Fig. 6. Cropped version of man image 2x magnification. From left to right: bi-cubic result, result of [14], result of the proposed method.

output of the MRF, without any post-processing, gives already reasonably good results. For example, all edges are sharp without “jaggy” artefacts which are visible in the result of bi-cubic interpolation. Back-projection further improves the result by eliminating artefacts and enhancing textures (e.g. texture on the roof). Finally, kernel regression only slightly smooths the image and can even be left out as a post-processing step in this case.

Table 1. RMSE and SSIM comparison of our method and bi-cubic interpolation result

Image	norm. RMSE		SSIM	
	Our	Bi-cubic	Our	Bi-cubic
Zebras	0.3589	0.3948	0.9097	0.9043
Skyscraper	0.2573	0.2789	0.9275	0.9163
Butterfly	0.1371	0.1484	0.9572	0.9564

We also compare our result with two state-of-the-art methods from [8] and [14]. In Fig. 6 we can see that the proposed method eliminates “jaggies” along the lines present in the results of reference methods, e.g. lines on the collar of the sweater. Our method also outperforms reference methods for higher magnification factor, as shown in Fig. 7. It manages to produce the sharpest lines without “jaggy” or “ghosting” artefacts that are present in the results of [8] and [14], while keeping the result visually pleasing. Both results were obtained with the input LR image of the small size. We believe that the difference would be even more significant for bigger input images.

In Table 1 we give quantitative results of a few images from Berkeley segmentation database[1]. We calculated the root mean square error (RMSE) and structure similarity index (SSIM) [15] between our super-resolution/bi-cubic interpolation result and ground truth. Our method produces smaller error and

¹ eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench

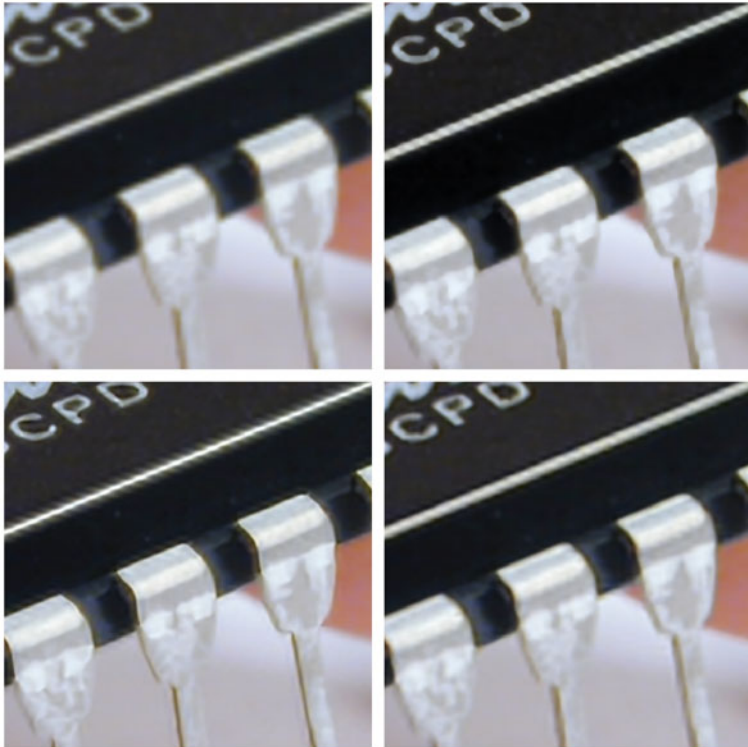


Fig. 7. Cropped version of chip image 4x magnification. Top-left: bi-cubic result. Top-right: result of [14]. Bottom-left: result of [8]. Bottom-right: result of the proposed method.

higher structure similarity score than bi-cubic interpolation. The quantitative improvement is, however, limited since the improvement is concentrated in edge regions, which represent small portion of the whole image.

4 Conclusion

In this paper we have presented a novel single-image super-resolution method based on MRF modelling. Unknown high-resolution image is modelled as a MRF whose nodes are overlapping high-resolution patches. Possible candidates for these nodes are found within only one level of the Gaussian pyramid of the input low-resolution image. To choose the best candidate in maximum a posteriori sense, we used our previously developed inference method called neighbourhood-consensus message passing, which makes this step fast and simple. Additionally, we performed back-projection and steering kernel regression to further improve the results. Results show that our method greatly outperforms standard techniques, while being visually better or comparable with state-of-the-art techniques.

References

1. Irani, M., Peleg, S.: Improving Resolution by Image Registration. *CVGIP: Graphical Model and Image Processing* 53(3), 231–239 (1991)
2. Farsiu, S., Robinson, D., Elad, M., Milanfar, P.: Fast and Robust Multi-Frame Super-Resolution. *IEEE Trans. Image Proc.* 13(10), 1327–1344 (2004)
3. Freeman, W.T., Pasztor, E.C., Carmichael, O.T.: Learning Low-Level Vision. *IJCV* 40(1), 24–47 (2000)
4. Baker, S., Kanade, T.: Limits on Super-Resolution and How to Break Them. *PAMI* 24(9), 1167–1183 (2002)
5. Tappen, M.F., Freeman, W.T.: Exploiting the Sparse Derivative Prior for Super-Resolution and Image Demosaicing. In: *IEEE Workshop on Stat. and Comp. Theories of Vision* (2003)
6. Kim, K., Kwon, Y.: Example-Based Learning for Single-Image Super-Resolution. In: Rigoll, G. (ed.) *DAGM 2008. LNCS*, vol. 5096, pp. 456–465. Springer, Heidelberg (2008)
7. Ebrahimi, M., Vrscay, E.: Solving the Inverse Problem of Image Zooming Using “Self-Examples”. In: Kamel, M.S., Campilho, A. (eds.) *ICIAR 2007. LNCS*, vol. 4633, pp. 117–130. Springer, Heidelberg (2007)
8. Glasner, D., Bagon, S., Irani, M.: Super-Resolution from a Single Image. In: *Int. Conf. on Comp. Vision, ICCV* (2009)
9. Luong, H., Ružić T., Pižurica, A., Philips, W.: Single Image Super-Resolution Using Sparsity Constraints and Non-Local Similarities at Multiple Resolution Scales. In: *SPIE* (2010)
10. Ružić T., Pižurica, A., Philips, W.: Neighbourhood-Consensus Message Passing and Its Potentials in Image Processing Applications. In: *SPIE Electronic Imaging* (2011)
11. Takeda, H., Farsiu, S., Milanfar, P.: Kernel Regression for Image Processing and Reconstruction. *IEEE Trans. Image Proc.* 16, 349–366 (2007)
12. Li, S.Z.: *Markov Random Field Modeling in Computer Vision*. Springer, Heidelberg (1995)
13. Yedidia, J.S., Freeman, W.T.: On the Optimality of Solutions of the Max-Product Belief-Propagation Algorithm in Arbitrary Graphs. *IEEE Trans. Inf. Theory* 47(2), 736–744 (2001)
14. Fattal, R.: Image Upsampling via Imposed Edge Statistics. In: *SIGGRAPH 2007* (2007)
15. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image Quality Assessment: From Error Visibility to Structural Similarity. *IEEE Trans. Image Proc.* 13(4), 600–612 (2004)

Background Images Generation Based on the Nelder-Mead Simplex Algorithm Using the Eigenbackground Model

Charles-Henri Quivy and Itsuo Kumazawa

Department of Information Processing, Tokyo Institute of Technology,
Nagatsuta-cho, Midori-ku, Yokohama, Kanagawa, Japan

Abstract. The Eigenbackground model is often stated to perform better than pixel-based methods when illumination variations occur. However, it has originally one demerit, that foreground objects must be small. This paper presents an original improvement of the Eigenbackground model, dealing with large and fast moving foreground objects. The method generates background images using the Nelder-Mead Simplex algorithm and a dynamic masking procedure. Experiments show that the proposed method performs as well as the state-of-the-art Eigenbackground improvements in the case of slowly moving objects, and achieves better results for quickly moving objects.

Keywords: Background subtraction, Eigenbackground model, Nelder-Mead Simplex algorithm.

1 Introduction

Thanks to the price decrease of the computing hardware, automatic video surveillance and monitoring applications are now widespread. Nevertheless, it is still necessary to analyse videos in the case of continuous surveillance applications, which requires time and expertise. Moreover, some applications induce unusual conditions that make difficult to run standard algorithms. For instance, the Taiwanese Ecogrid project [1] implements video monitoring of coral reef fishes. Since scenes are unconstrained, the size of foreground objects is unpredictable, as well as their speed.

Amongst statistical background models, subspace-learning models are stated to be designed for scenes illumination changes [2, 3], which include time-of-day and light-switching problems. The principle of subspace-learning methods is to compute a model of background images by reducing the dimensionality of the data. The Eigenbackground is one of the Reconstructive Subspace Learning (RSL) techniques [2]. It has two merits: background models are computed in an unsupervised way, and incremental subspace updating is possible (real-time applications). On the contrary, it presents two drawbacks: the size of foreground objects has to be small, and the model is not robust to outliers during the update

procedure. Numerous improvements have been proposed for solving the robustness problem [2], while only two methods deal with large foreground objects.

This paper focuses on monitoring applications that are similar to the Ecogrid Project. Even if the Eigenbackground algorithm is quite appropriate to such applications, it remains to improve the original method so that it deals with large and fast moving foreground objects.

The paper is structured as follows. Section 2 reviews the Eigenbackground model as well as the state-of-the-art improvements that deal with large foreground objects, and presents the shortcoming of the methods. Section 3 describes the method proposed to improve the generation of background images when large foreground objects are visible. Section 4 contains the experimental results. Lastly, Sect.5 is a discussion on the possible improvements of the proposed approach.

2 Related Works

The Eigenbackground algorithm [4] (EB) represents a background as the weighted sum of a mean background and p eigenvectors computed by applying Principal Component Analysis (PCA) on a training set of images. A training phase is first required to learn background variations. Then, during the running phase, the model is used to generate background images that are subtracted to incoming frames. Despite this paper does not deal with the model computation, it briefly explains this one in the following paragraph.

Let N background image vectors form a data set, $B = \{b_0, b_1, \dots, b_N\}$. It is required to compute the mean background of the data set, b_μ and the $N \times N$ covariance matrix C_b to conduct PCA. The covariance matrix is diagonalized using an eigenvalue decomposition: $C_b = \Phi_B \Lambda_b \Phi_B^{-1}$, where Φ_B is the eigenvector matrix and Λ_b the corresponding diagonal eigenvalues matrix. In order to keep significant background information and remove noise, only the eigenvectors corresponding to the p largest eigenvalues are kept [4], $p \leq N$. These eigenvectors (sometimes referred to as *variation modes* or *modes* in this paper) form the matrix Φ_b .

Thus, background image vectors are represented in a low dimensional space spanned by p eigenvectors (EB space), and can be reconstructed from this space using [1], with λ the parameter that controls background generation.

$$b = b_\mu + \Phi_b \cdot \lambda \quad (1)$$

Once the Eigenbackground model has been trained, it is possible to generate background images given the following procedure.

Let F_t be the input frame at time t . F_t is filtered by applying the projection step [2] onto the EB space, and the reconstruction step [1] from the subspace, which leads to the function defined by [3], with b_t the background computed from F_t .

$$\lambda_{F_t} = \Phi_b^{-1} \cdot (F_t - b_\mu) \quad (2)$$

$$b_t = b_\mu + \Phi_b \cdot \Phi_b^{-1} \cdot (F_t - b_\mu) \quad (3)$$

Eventually, the foreground image is computed by thresholding the absolute difference between F_t and b_t .

The EB model suffers from large foreground objects, as the projection onto the subspace depends on the amount of *noise* present in the frames. Two approaches have been proposed to overpass this limitation [2].

Zhifei et al. [5] introduced an algorithm that recursively improves the generation of background images in the case of large foreground object. This process iteratively detects foreground objects using the standard eigenbackground algorithm and replaces the pixels of the foreground objects with pixels of the mean background. Although this method provides better results than the original EB algorithm in the case of medium-sized foreground objects, it requires a plausible initial approximation of the background, which is not the case with very large foreground objects. It is referred to as REC (Recursive Error Compensation) in this paper.

Kawabata et al. [6] proposed a method that detects anomalous regions from dynamic scenes. At time t , the input frame F_t is combined with the previously computed background image: $E_t = \sigma \cdot F_t + (I - \sigma) \cdot b_{t-1}$, where σ is defined given the result of the previous frame background subtraction and I the identify. Then the background image b_t is computed using a trained EB $\{b_\mu, \Phi_b\}$ and E_t instead of F_t . The method performs well if objects are moving slowly, because the mask σ depends on the previous frame. It is referred to as AOD (Anomalous Objects Detection) in this paper.

The main limitation of the state-of-the-art methods stands in the projection of input images onto the EB space. On that basis, the novelty of this paper is to replace the filtering of input images F_t with a direct background images estimation, associated to a dynamic mask computation.

3 Proposed Method

The generation of background images from a trained EB $\{b_\mu, \Phi_b\}$ is based on the minimization of the objective function defined by (4), where λ is the vector that controls background generation, σ and K are scalars that control respectively the luminosity and the contrast of the generated images, and Σ is an occlusion mask defined as a grid of $N \times M$ regions corresponding to blocks of pixels on images. When the group of pixels (l, k) of the input image is occluded, $\Sigma(l, k) = 0$. Otherwise, $\Sigma(l, k) = 1$.

$$f(\lambda, \sigma, K) = \|\Sigma \cdot [F_t - (K \cdot (\sigma + b_\mu + \Phi_b \cdot \lambda))] \|_2 \quad (4)$$

The direct search Nelder-Mead simplex algorithm [7] stands as the basis of the proposed method, as it is stated to be particularly efficient in the first iterations [8]. Furthermore, it has already been proposed for Active Appearance Models fitting [9], which is related to the topic of this study. Nevertheless, the lack of convergence theory is a cause of concern.

At iteration i , the local euclidean distance $\epsilon^i(l, k)$ between F_t and the generated background is first computed for blocks (l, k) if $\Sigma(l, k) = 1$. Then the global euclidean distance E^i is computed from the local errors. During the minimization of (4), local errors $\epsilon^i(l, k)$ do not converge if the corresponding block of pixels (l, k) on F_t is a part of a foreground object. Such property is used to generate the occlusion mask after a few Nelder-Mead simplex iterations.

The generation of the background image b_t associated to F_t can be summarized as follows.

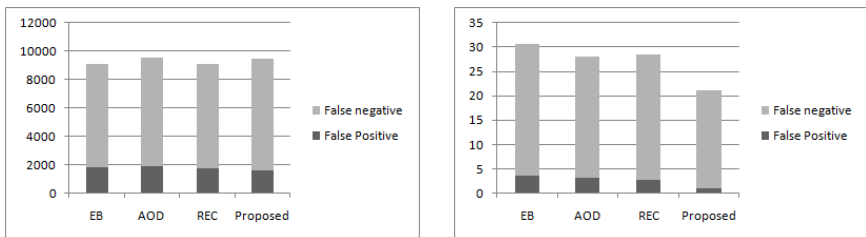
1. $\forall(l, k), \Sigma(l, k) = 1$
2. For iterations i from 1 to r , (4) is optimized and cumulated local errors $\epsilon_C^i(l, k)$ are computed
3. At iteration r , if $\epsilon_C^r(l, k) > T_h$ then $\Sigma(l, k) = 0$ where T_h is a threshold and (l, k) designates regions of pixels.
4. For iterations $r+1$ to $r+q$, (4) is optimized, $r+q$ representing the maximum number of iterations

The quality of generated background images mainly depends on the initialization of the Nelder-Mead simplex algorithm, but also on the threshold T_h used to generate the occlusion mask Σ . The benefit of this approach is that Σ only depends on the input frame F_t , contrarily to the AOD method.

4 Experiments

In order to evaluate the performance of the proposed method, the experiments are conducted on two data sets. The first data set (DS1) focuses on standard surveillance applications, i.e human tracking. The foreground objects of this data set are relatively small. On the other hand, the second data set (DS2) focuses on fish monitoring applications showing comparatively large and fast moving foreground objects.

DS1 is built from videos of the Wallflower data set, the IBM data set, and the PETS2001 database. The four videos of DS2 come from the Ecogrid project [1].



(a) Total FP/FN rates on the DS1 Wallflower images

(b) Mean FP/FN rates on DS2 videos (total of 366 images) in percentage

Fig. 1. False Positive (FP) and False Negative (FN) rates on images of DS1 and DS2

Table 1. Qualitative results on the DS1 images






















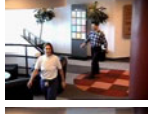

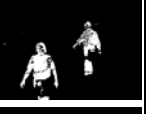


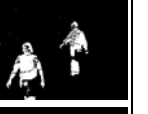
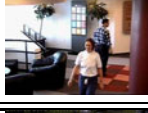










Original image	Ground truth	EB	REC	AOD	Proposed
					
					
					
					
					
					
					

Image resolution is 320×240 pixels for all videos except the wallflower data set, that is 160×120 . All the experiments follow the same conditions. The EB is trained with the first 200 images of video sequences. The dimensionality of the EB space is 10. The threshold for background subtraction is set to 50. Regarding the parameters of the REC method (defined in [5]), $\alpha = 1.2$, and $\epsilon = 5.0e^{-3}$. As for the proposed method, the cumulated error threshold T_h , defined in Sect. 3, is set to 50 for DS1, and to 15 for DS2. The maximum number of Nelder-Mead simplex algorithm is set to 20, and $r=6$. The size of the occlusion mask is 10×10 , i.e the size of each pixels region is 32×24 pixels, or 16×12 . The choice of the parameters' value is discussed in Sect. 5.

Table 1 shows the results of background subtraction on images of DS1. As the EB method is stated to perform poorly on the *Moved Object* and the *Bootstrapping* videos [3], these are removed from DS1. The segmentation results produced by REC, AOD and the proposed method are similar, with a small advantage to REC. To further evaluate the method qualitatively using DS1, a short video

Table 2. Short sequence of the Wallflower data set *Light Switch* case video (Frames 796 to 806)

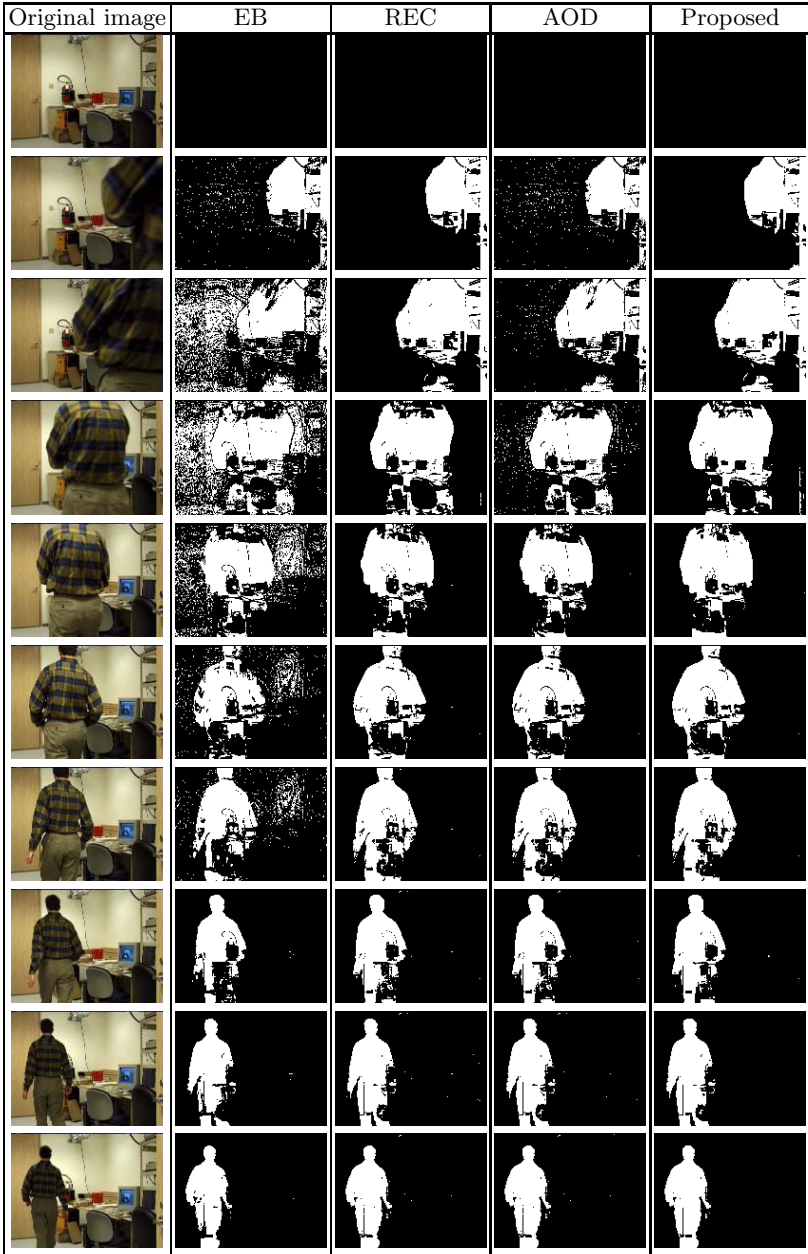
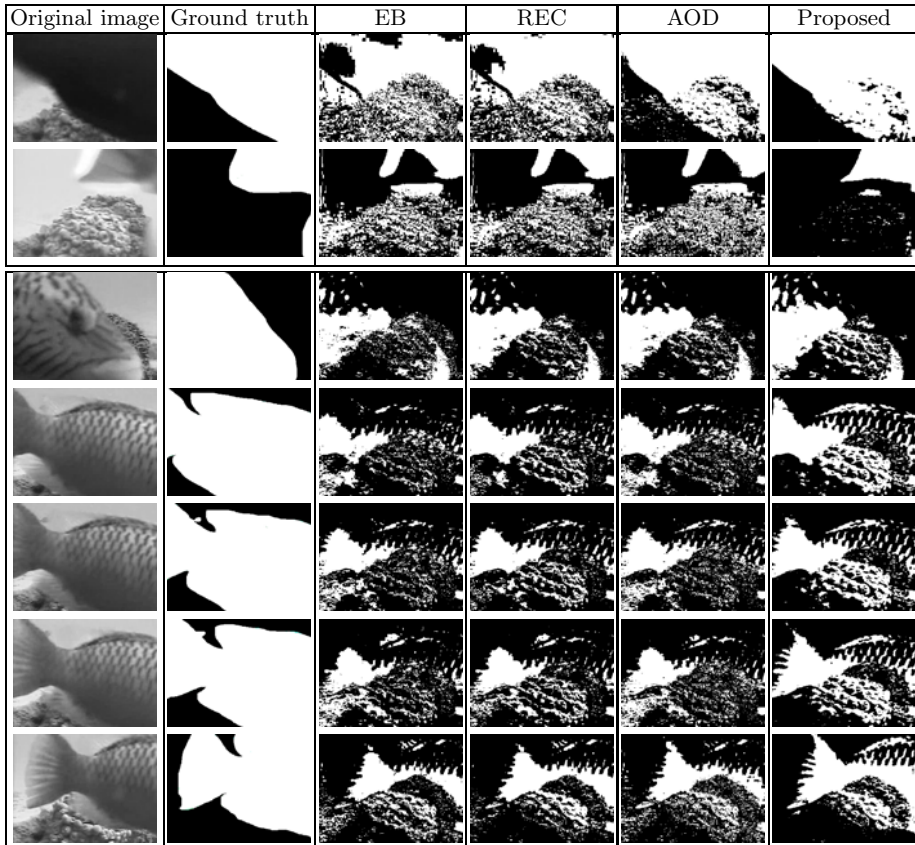


Table 3. Qualitative results on some DS2 images. The first two rows correspond to images of a video presenting a fish swimming from the left to the right (images 974 and 1008). The next rows present images 1184, 1238, 1243, 1248 and 1264 of another video.



sequence is extracted from the *Light Switch* case of the wallflower data base, as illustrated in table 2. On this test sequence, the REC method performs well, because the foreground object is relatively small (lower than 50%). The proposed method yields results similar to REC, while the AOD algorithm’s False Positive (FP) rate is the highest one (the foreground object is moving quickly). Some qualitative results on DS2 are presented in table 3. In general, the AOD method performs well, while the REC results show small improvements over EB results (because of the large size of the foreground object). On the other hand, the proposed method regularly outperforms both AOD and REC.

Regarding the quantitative evaluation of the method, Fig 1 exhibits the FP and False Negative (FN) rates for both DS1 and DS2. Results presented in Fig 1a are computed over only four images, but confirm the qualitative results. Figure 1b illustrates the mean FP rate computed over the four DS2 videos. In that case,

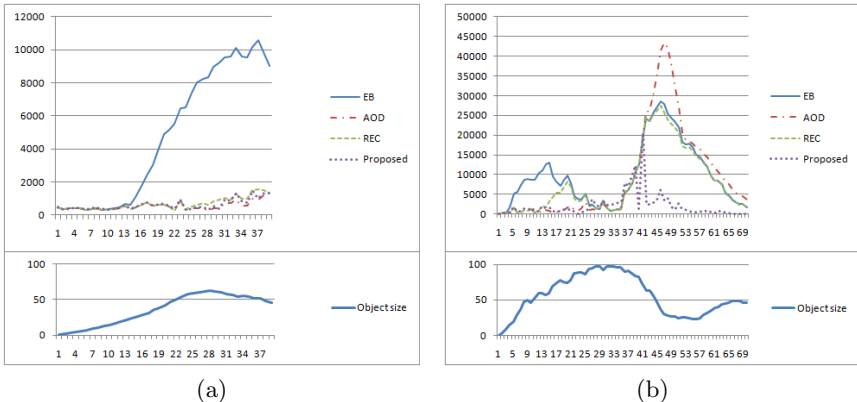


Fig. 2. Variation of the False Positive (FP) rate on two videos of DS2. The X-axis of the upper and the lower graphics represent the frames' number. The Y-axis of the upper and the lower graphics represent FP rates and the size of foreground objects in frame percentage, respectively.

the proposed method yields the lowest FN and FP rates. Figure 2a illustrates the variation of the FP rate over 39 frames of one video of DS2. The results are globally the same for AOD, REC and the proposed method. As expected, the standard EB algorithm performs poorly if the foreground object occupies more than 25% of the frame. Up to 60% of background occlusion by the foreground object, REC yields comparable results to AOD and the proposed method. From 60% to 100% (Fig 2b), AOD and the proposed method perform roughly the same. Nevertheless, when the size of the foreground object decreases suddenly (frame 35 and onwards, Fig 2b), the proposed method drastically outperforms the state-of-the-art techniques.

5 Discussion

This section briefly describes the influence of the parameters on the subtraction results. It mentions, as well, drawbacks of the proposed method.

In general, the parameters used for the evaluations of Sect. 4 are chosen experimentally, in order to produce qualitatively good results for both DS1 and DS2.

Although the number of variation modes p is usually set to keep between 95% and 98% of the training data set variance, p is arbitrarily set to 10 for all the experiments, which yields correct qualitative results. The choice of the threshold for background subtraction follows the same principle (results on the Wallflower data set images are qualitatively similar to related works' results).

Regarding the occlusion mask generation, a low r value may lead to the occlusion of parts of background, i.e wrong background image generation, whereas a high r value makes the proposed method perform the same as the EB algorithm. It is experimentally set to 6. The value of T_h depends on the value of r and the type of Nelder-Mead simplex algorithm initialization. As DS1 includes the *light switch* case video, the variation of parameter K (Eq 4) during initialization is larger than for DS2, which increases T_h value (respectively 50 and 15).

Finally, as the presence of a foreground object on the frame F_t makes the optimization algorithm converge towards a local minimum, the number of Nelder-Mead simplex algorithm iterations after the occlusion mask generation has to be greater than r , i.e. , $r + q \geq 2.r$.

The proposed method presents two major demerits. Firstly, it is based on an optimization algorithm, and the local minima problem becomes serious with the size of foreground objects. Secondly, despite the Nelder-Mead simplex algorithm is reportedly faster than other direct search algorithms, the proposed method is still slower than state-of-the-art techniques. In C language, one background image generation takes 0.817ms using openCV 2.0 on Core2Quad @ 2.4GHz for 320×240 pixels grey-level images. Dynamic down-sampling of images during optimization is a possible way of reducing the computational cost, and will be evaluated in a future work.

6 Conclusion

This paper presents an original method that replaces the projection/reconstruction step of the standard Eigenbackground algorithm with a direct background image generation. The experiments conducted on the two data sets DS1 and DS2 proved that the proposed method performs better than state-of-the-art approaches for large and fast moving objects. Otherwise, its performance is equivalent to such approaches on standard data sets. The future work is to focus on speeding-up the proposed technique, and to evaluate the method on a larger data set.

References

- [1] Ecogrid website (February 2010), <http://ecogrid.nchc.org.tw/sites.php?site=kt>
- [2] Bouwmans, T.: Subspace Learning for Background Modeling: A Survey. Recent Patent On Computer Science 2 (3), 223–234 (2009)
- [3] Toyama, K., Krumm, J., Brumitt, B., Meyers, B.: Wallflower: Principles and practice of background maintenance. 1, 255+ (1999)
- [4] Oliver, N.M., Rosario, B., Pentland, A.P.: A bayesian computer vision system for modeling human interactions. IEEE Transactions on Pattern Analysis and Machine Intelligence 22(8), 831–843 (2000)
- [5] Xu, Z., Shi, P., Gu, I.Y.-H.: An eigenbackground subtraction method using recursive error compensation. In: Zhuang, Y.-t., Yang, S.-Q., Rui, Y., He, Q. (eds.) PCM 2006. LNCS, vol. 4261, pp. 779–787. Springer, Heidelberg (2006)
- [6] Kawabata, S., Hiura, S., Sato, K.: Real-time detection of anomalous objects in dynamic scene. In: International Conference on Pattern Recognition, vol. 3, pp. 1171–1174 (2006)
- [7] Nelder, J.A., Mead, R.: A Simplex Method for Function Minimization. The Computer Journal 7(4), 308–313 (1965)
- [8] Lewis, R.M., Torczon, V., Trosset, M.W.: Direct search methods: then and now. Journal of Computational and Applied Mathematics 124(1-2), 191–207 (2000)
- [9] Aidarous, Y., Seguier, R.: Fast simplex optimization for active appearance model, pp. 106–117 (2009)

Phase Congruency Based Technique for the Removal of Rain from Video

Varun Santhaseelan and Vijayan K. Asari

University of Dayton, 300 College Park, Dayton, OH, USA
{santhaseelanv1,vijayan.asari}@notes.udayton.edu

Abstract. Rain is a complex dynamic noise that hampers feature detection and extraction from videos. The presence of rain streaks in a particular frame of video is completely random and cannot be predicted accurately. In this paper, a method based on phase congruency is proposed to remove rain from videos. This method makes use of the spatial, temporal and chromatic properties of the rain streaks in order to detect and remove them. The basic idea is that any pixel will not be covered by rain at all instances. Also, the presence of rain causes sharp changes in intensity at a particular pixel. The directional property of rain streaks also helps in the proper detection of rain affected pixels. The method provides good results in comparison with the existing methods for rain removal.

Keywords: Phase congruency, rain removal, alpha blending.

1 Introduction

Nowadays, video surveillance is an integral part of security applications. Outdoor video surveillance has helped in tackling serious law and order situations. It is only natural that with the increasing popularity of video surveillance equipment, the need for algorithms that improve video quality has also increased. One of the major challenges in video quality improvement when we consider outdoor vision systems is the effect of bad weather conditions on video.

Conditions that impede video quality include presence of haze, snow, fog, smoke, rain, hail, etc. Haze, smoke and fog can be considered as steady weather conditions and they fall in a different category of video enhancement. Rain and snow can be considered as dynamic weather conditions that change with every frame in the video. While rain is highly directional snow particles fall in completely random directions. This paper deals with the removal of rain from video.

The classification of weather into steady (haze, mist and fog) and dynamic (rain, snow and hail) weather was done by Garg and Nayar [1]. They developed models based on the physical and photometric properties of rain drops. They used these models to detect rain and to remove them from videos. The main assumption in that case was the uniform size of rain drops and the equal velocity of rain drops. The variation in depth was not taken into consideration. This became a problem while trying to remove rain from videos that contained heavy rain. Brewer and Liu also used the physical properties of rain drops to detect and remove rain from videos [2].

Garg and Nayar [3] also introduced an idea of changing camera parameters in order to reduce the effect of rain on the video. This method involved changing the camera parameters like F-number and exposure time individually or in tandem to reduce the effect of rain. The parameters were changed according to the nature of the scene. This method cannot be used in outdoor surveillance systems since manual adjustment of the camera parameters is not possible according to the weather conditions.

Park and Lee [5] came up with the idea of using a Kalman filter for the detection and removal of rain from videos. This method requires a periodic reset and cannot be adopted for videos taken from a moving camera. Barnum et al. [5] did a frequency space analysis of rain and snow affected videos. They modeled rain and snow in the frequency space based on the statistical properties of rain and snow streaks. Each rain streak was assumed to be a blurred Gaussian. The number of desired cycles to remove rain increases the number of frames to be used in the process.

Zhang et al. [6] used the spatio-temporal and chromatic properties of rain to remove rain from videos. Their idea was based on the fact that a pixel will not be covered by rain in every frame. They used an intensity histogram for each pixel constructed from all the frames in the video and used K-means clustering to differentiate between background pixels and rain affected pixels. This method works well except for the fact that all the frames in the video are used to construct the histogram.

The method proposed in this paper is along the lines of the idea used by Zhang et al. The proposed method uses phase congruency to detect candidate rain pixels. Since phase congruency is used, it is easier to incorporate the directionality property into the algorithm. The main advantage of this method in comparison to the method proposed by Zhang et al. is the fact that the number of frames used for detection and removal of rain affected pixels is minimal. Only the frames in the neighborhood are considered in this process.

The second section of this paper deals with the properties of rain streaks that appear on a video. This study has helped in the formulation of the algorithm. The third section explains in detail the steps involved in the algorithm and the feature extraction methods that have aided in rain detection and removal. Results and related discussion are included in the fourth section. A comparison with the existing methods is also provided in this section. The fifth section summarizes the findings in this paper and also discusses about the future work possible in this area.

2 Properties of Rain Streaks in Video

Most of the spatial, temporal and chromatic properties have been studied in detail by Zhang et al. These properties are utilized in this paper as well and are described briefly in this section.

2.1 Temporal Property

The human eye is able to see through rain mainly because all parts of the scene are not occluded by rain at all instances. As the depth of view increases, it becomes harder to distinguish between drops and the layer of rain appears as haze or mist [1].

This property holds true for occlusions due to rain in videos too. In this paper, we consider the removal of rain drops that can be distinguished separately in each frame. As the depth of view increases, the rain drops are not visible separately and the image enhancement problem becomes equivalent to haze removal. A close study of the intensity variation will show that the pixel intensity varies sharply when rain occludes a scene. This is illustrated in Fig. 1.

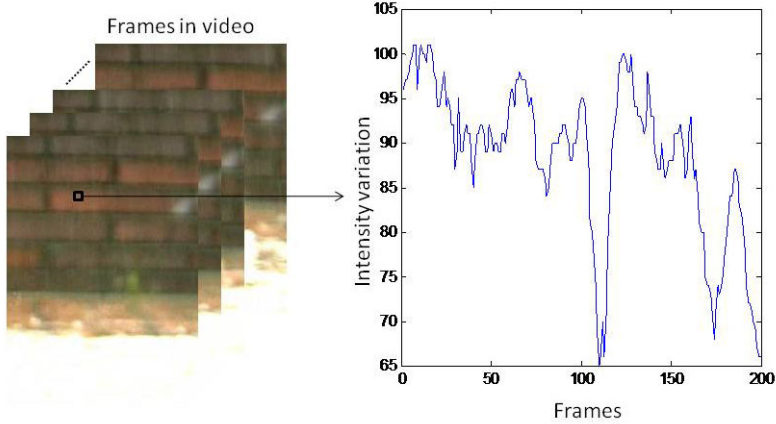


Fig. 1. Intensity variation for a pixel throughout a segment of the video containing heavy rain

The intensity variations plotted in Fig. 1 is for a video that contains heavy rain. It can be seen that the intensity tends to remain high if the density of rain is higher and therefore more frames will be required to compensate for the rain affected pixels. This is the case where considering one frame before and after the current frame becomes insufficient for rain removal.

2.2 Chromatic Property

While Garg and Nayar [1] showed that a rain drop refracts a wide range of light causing an increase in intensity at a particular pixel, Zhang et al. went ahead and showed that the change in levels for the individual color components of the pixel due to rain is proportional to its original intensity level. They showed that the standard deviation in each color component due to the presence of rain is almost the same.

2.3 Directional Property

Another observation that has been utilized by Garg and Nayar [1] is the directional property of rain in videos. If rain is present in a frame, all the rain streaks will be oriented in a single direction. They computed the correlation between neighboring pixels to detect rain affected pixels. This property is used in our proposed method while calculating phase congruency in a particular orientation.

3 Algorithm for Rain Detection and Removal

The proposed algorithm can be condensed into four steps as shown in Fig. 2.

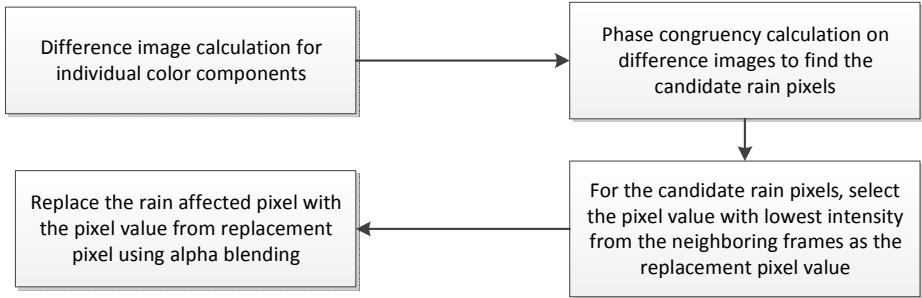


Fig. 2. Algorithm for rain detection and removal

3.1 Difference Image Calculation

The temporal property of rain described in the previous section indicates that there will be a positive change in intensity of a rain affected pixel. The chromatic property suggests that the standard deviation in all the three components will be the same when there is rain occluding a pixel. In this step, we compute the difference image of the current frame with respect to its neighbors. The difference image is computed for all the three color components separately. The neighboring frame is subtracted from the current frame. If the resultant value at a pixel is negative, it is clamped to zero. The presence of rain causes an increase in intensity. Therefore, only positive differences will be considered. For any pixel, if the standard deviation of the individual color components is different from each other, the pixel cannot be considered as rain-affected.

When differences of images are computed, the main criterion is the number of neighboring frames to be considered. As mentioned in section 2.1, if heavy rain is present, the number of frames to be considered will be more. In our case, we have used eight neighboring frames for the computation. This has resulted in good results with most of the rain removed from the video.

3.2 Applying Phase Congruency on the Difference Images

Phase information in the difference images are used to identify rain streaks in a particular frame. Phase congruency feature mapping gives an accurate measure of the variation in edges of rain streaks and is used in this paper.

3.2.1 Phase Congruency Features

The importance of phase information of an image is illustrated in Gonzalez and Woods [7]. The phase information in an image contains the essential details. When an image containing rain is considered, the rain streaks can be assumed to be the finer

details in the image. These fine details will be reflected in the phase changes of the image. This basic idea is the reason behind the inclusion of phase congruency feature detection as part of rain streak detection algorithm.

The principal reason that humans are able to visually recognize individual rain streaks in a particular frame is because there is a step change in intensity along the edge of the rain streak. Phase congruency (PC) is a feature detection mechanism that recognizes those edges and is invariant to illumination and contrast. The key observation that led to the development of phase congruency algorithm is that the Fourier components of an image are maximal in phase where there are edges or lines. Features are identified according to the extent to which the Fourier components are in phase.

The PC computation method adopted in this paper was proposed by Peter Kovessi [8]. His method was based on the local energy model developed by Morrone and Owens [9]. They observed that the point of strong phase congruency corresponds to a point of maximum energy. Let $I(x)$ be an input periodic signal defined in $[-\pi, \pi]$. $F(x)$ is the signal ($I(x)$) with no DC component and $H(x)$ is the Hilbert Transform of $F(x)$ which is a 90° phase shifted version of $F(x)$. The local energy, $E(x)$ can then be computed from $F(x)$ and its Hilbert Transform as in (1).

$$E(x) = \sqrt{F^2(x) + H^2(x)} \quad (1)$$

It has been shown in earlier research [10] that the energy is equal to the product of phase congruency and the sum of Fourier amplitudes as in (2).

$$E(x) = PC(x) \sum_n A_n \quad (2)$$

Therefore the peaks in phase congruency correspond to the peaks in the energy function. Equation (2) also shows that the phase congruency measure is independent of the overall magnitude of the signal, thus making the feature invariant to changes in illumination and contrast. The components, $F(x)$ and $H(x)$ are computed by the convolution of the signal with a quadrature pair of filters. Logarithmic Gabor filters are used in this case. Consider $I(x)$ as an input signal and M_n^e and M_n^o are the even symmetric and odd symmetric components of the log Gabor function at a particular scale, n . Then the amplitude and phase for the input signal in the transformed domain is obtained as in (3) and (4) where $o_n(x)$ and $e_n(x)$ are the responses for each quadrature pair of filters as given in (5).

$$A_n = \sqrt{e_n^2(x) + o_n^2(x)} \quad (3)$$

$$\phi_n(x) = \tan^{-1}(o_n(x)/e_n(x)) \quad (4)$$

$$[e_n(x), o_n(x)] = [I(x) * M_n^e, I(x) * M_n^o] \quad (5)$$

The values for $F(x)$ and $H(x)$ can be computed as shown in (6) and (7).

$$F(x) = \sum_n e_n(x) \quad (6)$$

$$H(x) = \sum_n o_n(x) \quad (7)$$

When the Fourier components are very small, the problem of computing phase congruency becomes ill-conditioned. This problem is solved by adding a small constant to the sum of Fourier components as shown in (8).

$$PC(x) = \frac{E(x)}{\varepsilon + \sum_n A_n} \quad (8)$$

Equation (8) is the final equation for solving phase congruency. This equation can be applied to a two dimensional signal like an image for various orientations. In this paper, the analysis is to be done on an image.

For an image, the first step is to convolve the image with a bank of two dimensional log Gabor filters. The filter has a transfer function as shown in (9).

$$G(w) = e^{(-\log(w/w_0)^2)/(2\log(k/w_0)^2)} \quad (9)$$

where w_0 is the filter's center frequency and k/w_0 is kept constant for various w_0 . The cross-section of the transfer function of the filter can be represented as in (10).

$$G(\theta) = e^{-(\theta-\theta_0)^2/(2\sigma_\theta^2)} \quad (10)$$

where θ_0 represents the orientation of the filter and σ_θ is the standard deviation of the Gaussian spreading function in the angular direction. As in equation (5) the even symmetric and odd symmetric components at a particular scale and orientation can be computed as shown in (11).

$$[e_{no}(x, y), o_{no}(x, y)] = [I(x, y) * M_{no}^e, I(x, y) * M_{no}^o] \quad (11)$$

The amplitude of the response at a particular scale and orientation can be computed as in (12), and the calculation of phase congruency for an image is as shown in (13).

$$A_{n0} = \sqrt{e_{no}^2(x, y) + o_{no}^2(x, y)} \quad (12)$$

$$PC(x, y) = \frac{\sum_o \sqrt{(\sum_n e_{no}(x, y))^2 + (\sum_n o_{no}(x, y))^2}}{\varepsilon + \sum_o \sum_n A_{no}(x, y)} \quad (13)$$

In this paper, all the orientations are not considered when phase congruency features are computed. This is because of the directional property of rain streaks. The rain drops always fall towards the ground and the variation in orientation is minimal. This fact helps in discarding most of the orientations. The calculation of difference images and phase congruency features are illustrated in Fig. 3.

3.3 Background Pixel Search

After applying phase congruency, only the candidate pixels (rain affected pixels) with intensity variations in neighboring frames remain in the processed image. The next step is to eliminate the false positives which may have occurred due to the presence of external noises. If a pixel is detected as a candidate rain pixel in all the phase congruency images of the difference images, it is very likely that it happened due to noise. These pixels are eliminated from the group of candidate rain pixels.

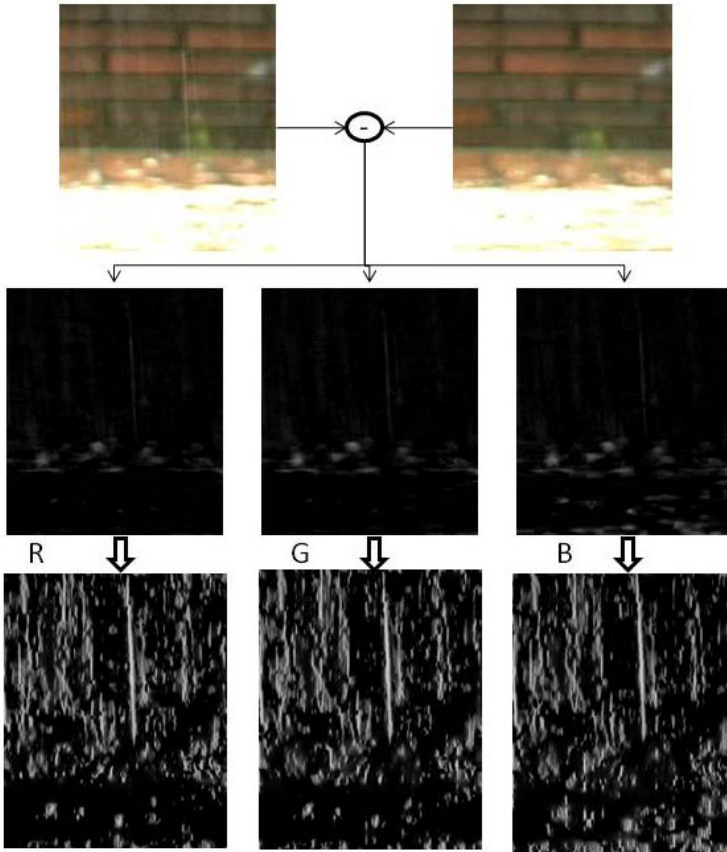


Fig. 3. The computation of difference image and the image with phase congruency features for R, G and B components

The next step is to find out the background intensity levels of the rain affected pixels. A search is performed on the neighboring frames. The pixel value that has the lowest intensity levels within the neighbors is selected as the background intensity of the rain affected pixel.

3.4 Compensate for Rain Affected Pixels

Garg and Nayar [1] used the average of the pixel intensities in neighboring two frames to compute the intensity value for the pixel to be replaced. This method fails when the pixel is affected by rain continuously. The method by Zhang et al. gave better results. They used alpha-blending to calculate the intensity value for the rain affected pixel as shown in (14).

$$C = \alpha C_b + (1 - \alpha) C_r \quad (14)$$

The new color is denoted as C , the background color is denoted as C_b and the color of the rain-affected pixel is denoted as C_r .

4 Results and Discussion

The results shown in Fig. 4 show that phase congruency features can be used in differentiating rain streaks from the original scene with the help of the spatio-temporal and chromatic properties of rain.

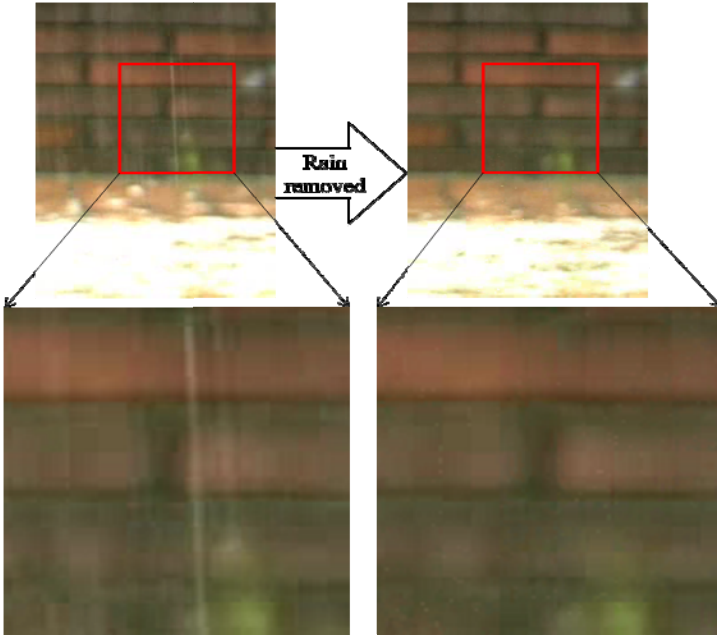


Fig. 4. The rain streaks in the frame shown on the left side have been removed and the resultant image is shown on the right side. Please refer the following link for the complete video: <http://visionlab.udayton.edu/research/rain.php>.

A performance comparison of the proposed algorithm is illustrated in Fig. 5. In comparison with the algorithm presented in [6], it is observed that the dynamic nature of the scene is preserved more in our method. For example, the intensity variations caused on the water puddles due to water drops are preserved more in Fig. 5(c). This is because our algorithm used comparatively much lesser number of frames (eight frames in the present experiment) for the removal of rain.

The important factor in our algorithm that affects the quality of the output video is the number of neighboring frames that are considered. It has been observed that increasing the number of frames will increase the quality of video, especially when the aim is to remove heavy rain. This increase in quality comes at the expense of loss of preservation of motion of objects in the frame. The effect of increasing the number of frames is illustrated in Fig. 6. While one trial used six neighboring frames, twelve

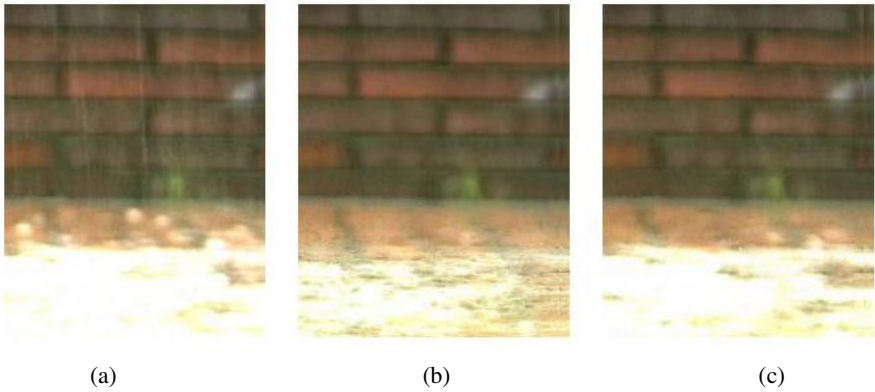


Fig. 5. This figure compares the result between our method and the method by Zhang et al. [6]. (a) The original frame; (b) Rain removed by the method in [6]. (c) Rain removed by our method.

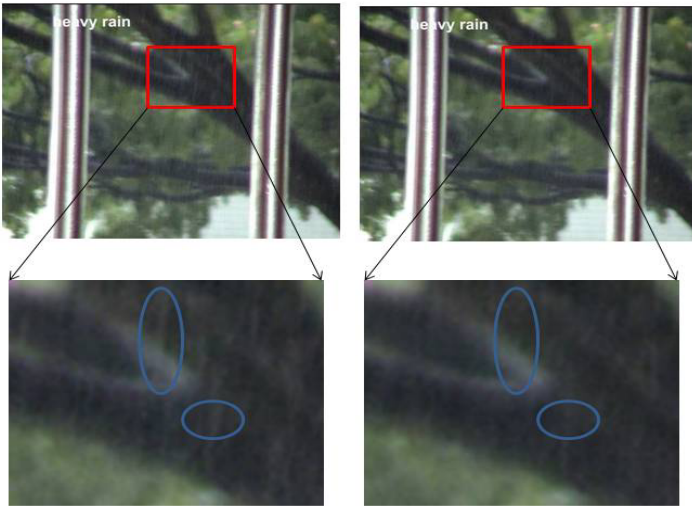


Fig. 6. The image on left is a video frame from which rain was removed using six neighboring frames and the image on the right utilized twelve neighboring frames. The presence of extra streaks in image shown on the left are highlighted.

frames were used for the second trial. It was observed that the addition of more frames for compensation reduced the number of blurred streaks in every frame.

5 Conclusion

A new method based on phase congruency features was used to detect and remove rain from videos. The method was formulated based on the temporal, spatial and

chromatic properties of rain streaks in video. In comparison with the method of Zhang et al., it has been found that our method provides results of the same quality with lesser number of frames. It was also observed that the slight movements of objects in the video are captured better in our method.

This paper dealt with removal of rain from videos that did not have any camera movement. One way to deal with such a scenario is to stabilize the video [11] before applying the algorithm for rain removal as done by Zhang et al. Another area for future improvement is to tackle the problem of moving objects in the foreground of the rain as well as in the rain. In such cases, the aim will be to estimate the rain component in video from lesser number of frames.

References

1. Garg, K., Nayar, S.: Vision and rain. *International Journal of Computer Vision* 75, 3–27 (2007)
2. Brewer, N., Liu, N.: Using the shape characteristics of rain to identify and remove rain from video. In: da Vitoria Lobo, N., Kasparis, T., Roli, F., Kwok, J.T., Georgiopoulos, M., Anagnostopoulos, G.C., Loog, M. (eds.) *S+SSPR 2008*. LNCS, vol. 5342, pp. 451–458. Springer, Heidelberg (2008)
3. Garg, K., Nayar, S.K.: When does a camera see rain? In: *International Conference on Computer Vision 2005*, pp. 1067–1074 (October 2005)
4. Park, W.J., Lee, K.H.: Rain removal using Kalman filter in video. In: *International Conference on Smart Manufacturing Application*, pp. 494–497 (April 2008)
5. Barnum, P., Kanade, T., Narasimhan, S.: Spatio-temporal frequency analysis for removing rain and snow from videos. In: *Workshop on Photometric Analysis For Computer Vision (2007)*
6. Zhang, X., Li, H., Qi, Y., Leow, W.K., Ng, T.K.: Rain removal in video by combining temporal and chromatic properties. In: *IEEE International Conference on Multimedia and Expo 2006*, pp. 461–464 (July 2006)
7. Gonzalez, R.C., Woods, R.E.: *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc., Boston (1992)
8. Kovsi, P.: Image features from Phase Congruency. *Visere: Journal of Computer Vision Research* 1(3) (Summer 1999)
9. Morrone, M.C., Owens, R.A.: Feature detection from local energy. *Pattern Recognition Letters* 6, 303–313 (1987)
10. Venkatesh, S., Owens, R.A.: An energy feature detection scheme. In: *The International Conference on Image Processing*, pp. 553–557 (1989)
11. Matsushita, Y., Ofek, E., Tang, X., Shum, H.Y.: Full-frame video stabilization with motion inpainting. In: *Proceedings of CVPR 2005*, vol. 1, pp. 50–57 (2005)

A Flexible Framework for Local Phase Coherence Computation

Rania Hassen, Zhou Wang, and Magdy Salama

Department of Electrical and Computer Engineering, University of Waterloo,
200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1
{raniahassen,zhouwang}@ieee.org, M.Salama@ece.uwaterloo.ca

Abstract. Local phase coherence (LPC) is a recently discovered property that reveals the phase relationship in the vicinity of distinctive features between neighboring complex filter coefficients in the scale-space. It has demonstrated good potentials in a number of image processing and computer vision applications, including image registration, fusion and sharpness evaluation. Existing LPC computation method is restricted to be applied to three coefficients spread in three scales in dyadic scale-space. Here we propose a flexible framework that allows for LPC computation with arbitrary selections in the number of coefficients, scales, as well as the scale ratios between them. In particular, we formulate local phase prediction as an optimization problem, where the object function computes the closeness between true local phase and the predicted phase by LPC. The proposed method not only facilitates flexible and reliable computation of LPC, but also demonstrates strong robustness in the presence of noise. The groundwork laid here broadens the potentials of LPC in future applications.

Keywords: local phase coherence, scale-space, complex wavelet coefficients, feature detection.

1 Introduction

Phase information plays a crucial role in preserving important structural features in various types of signals, including 1D (e.g., speech), 2D (e.g., still images) and 3D (e.g., video or volume data) signals. For example, if the Fourier transform domain amplitude and phase spectra of two images are interchanged, the resulting hybrid image is recognized from which the phase spectrum is taken [1]. In understanding the structures of natural images, however, *global* Fourier phase may not be the best option, because natural images tend to be non-stationary, with different sizes and shapes of smooth or periodic regions, and distinctive features (such as edges and lines) between them. Furthermore, physiological studies suggest that many neurons located in the visual cortex are best models as filters localized in space, frequency and orientation. As a result, *local* phase is a more plausible quantity in cortical encoding and processing. In terms of image processing and understanding, it is also a better tool in describing the structures of natural images.

In the pioneering work in studying the relation between congruence of local phases [2,3], a local energy model was introduced which postulates that in a waveform which have unique perceptual significance as “lines” and “edges”, the Fourier components come into phase with each other at these feature points. Based on this observation, it was suggested that the visual system could locate features of interest by searching for maxima of local energy points, and identify the feature type by evaluating the value of arrival or local phase at that point [2,3]. Almost all work thereafter concentrated on finding points of maximal phase congruency by looking for maxima in local energy. In [4], a direct measure of phase congruency was proposed, where a phase congruency measure is computed as a dimensionless quantity that is invariant to changes in image brightness or contrast and thus provides an absolute measure of the significance of feature points. Through the use of wavelets, an extension from 1D to 2D phase congruency calculation is also developed [4]. Local phase based method has also been employed in a number of computer vision and image processing problems, including estimation of image disparity [5] and motion [6,7], description of image texture [8], recognition of persons using iris patterns [9], and video quality assessment [10].

In [11], the local phase structures at distinctive features were examined in more depth. The local phase coherence (LPC) relationship was first discovered, which not only predicts the alignment of phases across scales *at* the location of features (as found in earlier work), but also describes the full structure of local phase pattern in scale-space in the *vicinity* of feature location. It was suggested in [11] that the LPC relation could lead to a new theory in the perception of blur, and may have deeper implications on how the visual system could “see beyond the Nyquist rate”. Since the introduction of LPC, it has been found to be useful in a number of applications, including image registration [12], fusion [13] and sharpness evaluation [14]. One common limitation in all existing applications is that the LPC can only be computed with 3 coefficients spread in 3 scales in dyadic scale-space. This restricts its application, especially when one would like to have a closer examination of LPC relationship in the scale-space where smaller (and fractional) scale ratios are desired. The purpose of the current study is to develop a flexible methodology in computing LPC and thus extends its potentials in real applications.

2 Local Phase Coherence and Computation

2.1 Local Phase Coherence

The concept of LPC is built upon complex wavelet analysis tools that provide localized magnitude and phase information in multi-scales. Given a signal $f(x)$ localized near the position x_0 , where $f(x) = f_0(x - x_0)$, a general complex wavelet transform may be written as:

$$F(s, p) = \int_{-\infty}^{\infty} f(x) w_{s,p}^*(x) dx = \left[f(x) * \frac{1}{\sqrt{s}} g\left(\frac{x}{s}\right) e^{j\omega_c x/s} \right]_{x=p}, \quad (1)$$

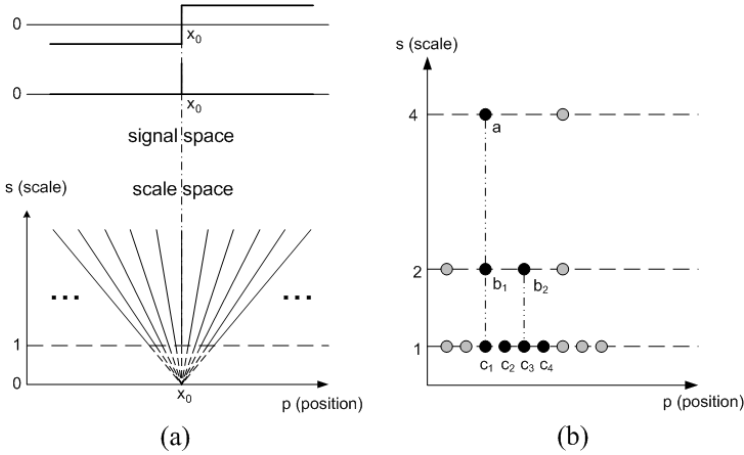


Fig. 1. (a) Local phase coherence structure near localized feature. (b) An example of 1D sampling grid in scale-space.

where $s \in R^+$ is the scale factor, $p \in R$ is the translation factor, and the family of wavelets are derived from the mother wavelet $w(x) = g(x)e^{j\omega_c x}$ by

$$w_{s,p}(x) = \frac{1}{\sqrt{s}} w\left(\frac{x-p}{s}\right) = \frac{1}{\sqrt{s}} g\left(\frac{x-p}{s}\right) e^{j\omega_c(x-p)/s}, \quad (2)$$

where ω_c is the center frequency of the modulated band-pass filter, and $g(x)$ is a slowly varying and symmetric envelop function. Here the wavelet is considered general because we do not specify $g(x)$, which has many different options but the theory derived here applies to all.

Using the convolution theorem, and the shifting and scaling properties of the Fourier transform, we can derive:

$$F(s,p) = \frac{1}{2\pi\sqrt{s}} \int_{-\infty}^{\infty} F_0\left(\frac{\omega}{s}\right) G(\omega - \omega_c) e^{j\omega(p-x_0)/s} d\omega, \quad (3)$$

where $F(\omega)$, $F_0(\omega)$ and $G(\omega)$ are the Fourier transforms of $f(x)$, $f_0(x)$ and $g(x)$, respectively. The phase of $F(s,p)$ depends on the nature of $F_0(\omega)$. If $F_0(\omega)$ is scale invariant, meaning that $F_0(\omega/s) = K(s)F_0(\omega)$, where $K(s)$ is a real function of only s , but independent of ω , then it is not hard to find that:

$$F(s,p) = \frac{K(s)}{\sqrt{s}} F\left(1, x_0 + \frac{p-x_0}{s}\right). \quad (4)$$

Since both $K(s)$ and s are real, we obtain the following phase relationship of $F(s,p)$:

$$\Phi(F(s,p)) = \Phi\left(F\left(1, x_0 + \frac{p-x_0}{s}\right)\right). \quad (5)$$

This result indicates that there is a strong phase coherence relationship across scale and space, where equal phase contours in the (s,p) plane form straight lines

that converge exactly at the location of the feature x_0 , as illustrated in Fig. [11\(a\)](#). These straight lines are defined by $x_0 + (p - x_0)/s = C$, where C is a constant. Note that this result is based on the assumption that f_0 is a scale invariant signal, which turns out to be true for distinctive features (such as an impulse or a step edge in a 1D signal, or an edge or line in a 2D image). Therefore, LPC measurement can be used to detect distinctive features in a signal.

2.2 Computation of Local Phase Coherence

If the LPC relationship is satisfied at a spatial location, then the phase of a wavelet coefficient may be predicted by the phases of its neighboring coefficients in the scale-space. Conversely, the prediction accuracy could be used as a measure of the strength of LPC. This approach was first employed in [11](#). An example is shown in Fig. [11\(b\)](#), where the finest scale coefficients c_i for $i = 1, 2, 3, 4$ can be predicted from their coarser scale neighbors a , b_1 and b_2 . For example,

$$\hat{\Phi}(c_1) = -2\Phi(a) + 3\Phi(b_1). \quad (6)$$

Although such prediction can lead to useful measures of the strength of LPC and has been successfully used in several applications [11,12,13,14](#), it is limited to grouping three coefficients at a time that are separated into three scales with fixed scale ratio of 2 between successive scales, as exemplified in Fig. [11\(b\)](#). Here we propose a novel framework that allows for more flexibility in the computation of LPC. Let us consider a group of N coefficients a_i for $i = 1, \dots, N$, each of which is a sample of $F(s, p)$ at (s_i, p_i) , i.e., $a_i = F(s_i, p_i)$. If the LPC relationship is satisfied, then we should be able to best predict the phases of these coefficients, i.e., the error between the predicted and true phases should be minimized. The simplest form of an error function is the mean squared error

$$E_1 = \frac{1}{N} \sum_{i=1}^N \left(\Phi(a_i) - \hat{\Phi}(a_i) \right)^2. \quad (7)$$

Note that for distinctive features such as a line or a step edge, the phase pattern in the scale-space can be approximated using a functional form. For example, in the case of a step edge $f_0(x) = K[u(x) - \frac{1}{2}]$, we have:

$$\hat{\Phi}(F(s, p)) \approx \frac{w_c(p - x_0)}{s} - \frac{\pi}{2} + n_1\pi, \quad (8)$$

where the constant term ($-\frac{\pi}{2}$ here) depends on feature type, for another example, in the case $f_0(x) = K\delta(x)$, the constant is 0. Assuming that a set of coefficients are aligned at the same position p but across consecutive scales $s_i = 1, r, r^2, \dots, r^{N-1}$, where r is the scale ratio between successive scales that may be any fractional number greater than 1. Further, we simplify the phase prediction expression by denoting $Q_p = w_c(p - x_0)$. Then the problem of solving for best phase prediction is converted to

$$Q_p^{(opt)} = \arg \min_{Q_p} E_1. \quad (9)$$

This can be solved by setting $\partial E_1/\partial Q_p = 0$ and solve for Q_p , which leads to the following closed-form solution

$$Q_p^{(opt)} = \frac{\sum_{i=1}^N r^{(n-i)} \left(\Phi(a_i) + \frac{\pi}{2} \right)}{\sum_{i=1}^N r^{2(i-1)}}. \quad (10)$$

In the case that the coefficients in scale-space are located at more than one position, say M , then we can solve for a series of Q_p values $Q_{p_1}, Q_{p_2}, \dots, Q_{p_M}$ using similar approaches. After computing all Q_p values, we will be able to calculate the predicted phases for all coefficients. We can then define an LPC measure as

$$PC_1 = \frac{\Re \left\{ \prod_i a_i e^{-j\hat{\Phi}(a_i)} \right\}}{\prod_i |a_i| + C_1} = \frac{\Re \left\{ \prod_i |a_i| e^{j[\Phi(a_i) - \hat{\Phi}(a_i)]} \right\}}{\prod_i |a_i| + C_1}, \quad (11)$$

where the numerator is the real part of the phase prediction error in the complex plane weighted by the coefficient magnitude, so that the coefficients with higher magnitudes are given more importance. The result is normalized by the magnitude of the coefficients. C_1 is a small positive constant in order to stabilize the measurement when the signal is close to flat, in which case the coefficients have near zero magnitudes. This measurement states that if the predicted phases are very close to the actual phases and the signal is significant (such that their magnitude is significantly larger than the constant C_1), then we achieve good LPC with a PC_1 value close to 1. At the other extreme, if the predicted phases are perpendicular to the true phases, then the value of PC_1 will be close to 0.

Although the above method and solution is elegant in the sense that it offers closed-form analytical solution, it does not give us satisfactory results in our experiments. This may be partially due to the 2π wrap-around effect of angular variables (for which direct least square error function is deemed not appropriate). It may also be because the constant terms in the phase prediction form (e.g., Eq. (8)) varies for different types of features. For example, the $-\frac{\pi}{2}$ term in Eq. (8) would be $+\frac{\pi}{2}$ for a step edge $f_0(x) = K[\frac{1}{2} - u(x)]$ and 0 for an impulse $f_0(x) = K\delta(x)$.

To overcome the above problems, we define a new error energy function between the true and predicted phases as follows

$$E_2 = \left[1 - \frac{1}{N} \sum_{i=1}^N \cos \left(4\Phi(a_i) - 4\hat{\Phi}(a_i) \right) \right]^2. \quad (12)$$

The trick here is to multiply the angles by a factor of 4. This eliminates the ambiguities between the types of features because all the feature-dependent phase constants (such as the $-\frac{\pi}{2}$ term in Eq. (8)) are raised to a multiplier of 2π . In addition, the use of the cosine function in Eq. (12) avoids the 2π wrap-around effect of angular variables. Notice that when the phase prediction is in effect, the difference $4\Phi(a_i) - 4\hat{\Phi}(a_i)$ will be either close to 0 or a multiplier of 2π , and thus

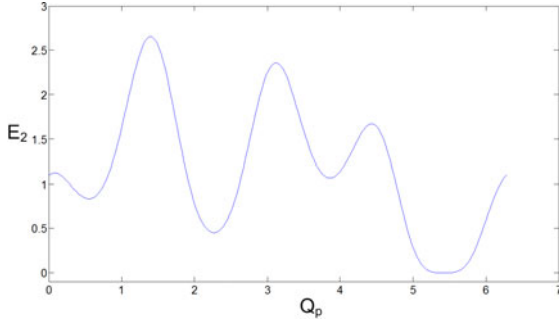


Fig. 2. An example of the search space of E_2 against Q_p

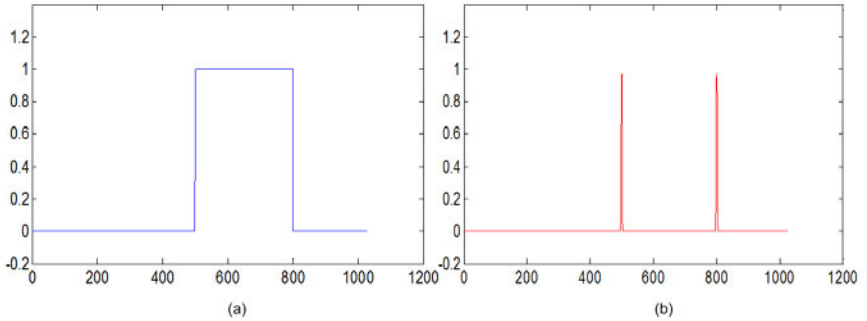


Fig. 3. (a) Original signal; (b) calculated LPC using Eq. (13)

the cosine of it will be close to 1. Consequently, the total error energy function E_2 will be close to 0.

Although the definition of E_2 has many good properties, it is a difficult function to optimize. For example, finding Q_p using a closed-form solution like Eq. (10) is difficult. Indeed, E_2 could be a fairly complicated function. An example is shown in Fig. 2, where the function of E_2 with respect to Q_p is smooth but has many local minima. In our implementation, we use an iterative numerical method to minimize the function, where the full search range is divided into 8 equally spaced segments, each associated with a different initial point at the center of the segment as the initial guess in the iteration. This results in multiple local minima, and then the global minimum is obtained by picking the lowest local minima. Finally, the LPC is computed by

$$PC_2 = \frac{\Re \left\{ \prod_i (a_i)^4 e^{-j4\hat{\Phi}(a_i)} \right\}}{\prod_i |(a_j)^4| + C_2} = \frac{\Re \left\{ \prod_i |(a_i)|^4 e^{j4(\hat{\Phi}(a_i) - \hat{\Phi}(a_i))} \right\}}{\prod_i |(a_j)|^4 + C_2}. \quad (13)$$

Similar to Eq. (11), C_2 is a positive stabilizing constant, and this is an energy weighted phase consistency measure, where the maximal value is achieved if all

phase predictions are perfect. Fig. 3(a) shows a simulated signal with ideal step edges, and Fig. 3(b) gives the LPC computation result using Eq. (13). It can be seen high PC_2 values are achieved (high peaks) at the step edges.

3 Simulations

In this section we will present several experiments meant to gauge the performance and robustness of the proposed technique for LPC computation. Although the experiments were carried out in 1D (which helps us better visualize the performance of the algorithm), similar techniques can also be applied to 2D or higher dimensional signals.

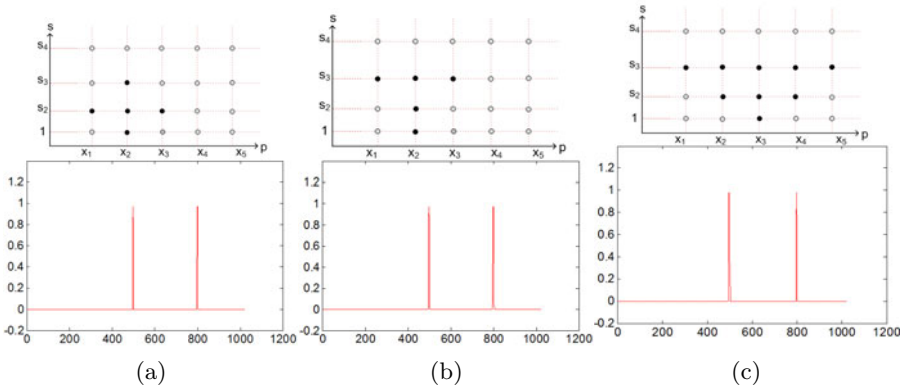


Fig. 4. LPC computation by grouping local coefficients in three different ways

The first experiment aims to demonstrate the flexibility of our framework in picking arbitrary group of neighboring coefficients in LPC computation. The upper figures in Fig. 4 show three different selections of complex wavelet coefficients in the scale-space, where the coefficients spread in three scales and up to five spatial locations. The scale ratios between successive scales are fixed at 2. The lower figures show the LPC measure PC_2 computed as a function of space for the signal in Fig. 3(a). Despite the quite different coefficient grouping, it can be observed that the resulting phase coherence functions are approximately the same. This result suggests that in practice, LPC can be computed in the complex wavelet transform domain with any coefficient setup, and may also be useful in the applications where only partial information of the local phase measurement is available.

The second experiment demonstrates the flexibility in picking scale ratios between successive scales. Most existing wavelet transforms were designed in dyadic scale-space, i. e., the scale ratio between successive scales is fixed at 2. From the derivations in the last section, this should not be a necessary condition in the computation of LPC. The scale ratio can be any other fractional number greater than 1. Even further, the scale ratio does not have to be the same between

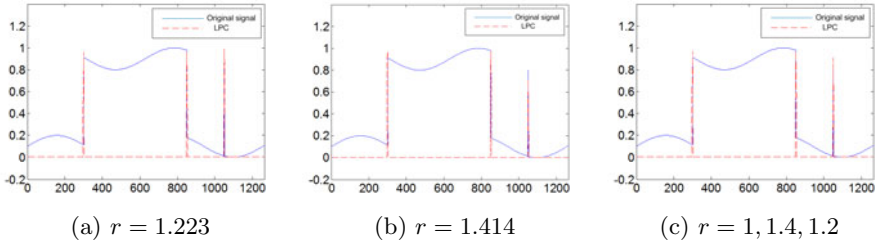


Fig. 5. LPC computation for a signal by using fractional scale ratios between coefficients. (a) Fixed scale ratio of $r = 1.223$ between 3 consecutive scales; (b) Fixed scale ratio of $r = 1.414$ between 3 consecutive scales; (c) Varying scale ratio $r = 1.4$ between the first 2 scales and $r = 1.2$ between the second and third scales.

Scales 1 to 2 and Scales 2 to 3. Figure 5 shows the resulted phase coherence using different setup of scale ratios. In the first two example, the scale ratios are fixed across three scales but are fractional numbers of $r = 1.223$ and $r = 1.414$, respectively. In the third example, the scale ratio is varying between the first two scales $r = 1.4$ and the last two scales $r = 1.2$. In all three cases, the resulting PC_2 functions are almost the same when applied to the same signal. This is a useful feature in practical applications because real world signals often contain mixtures of many distinctive features, and thus local measurement up to coarse scales often suffers from interference from nearby features. If the scale ratios can be fractional (preferably less than 2), then we will be able to carry out closer scale-space analysis of local features and avoid interference from nearby features.

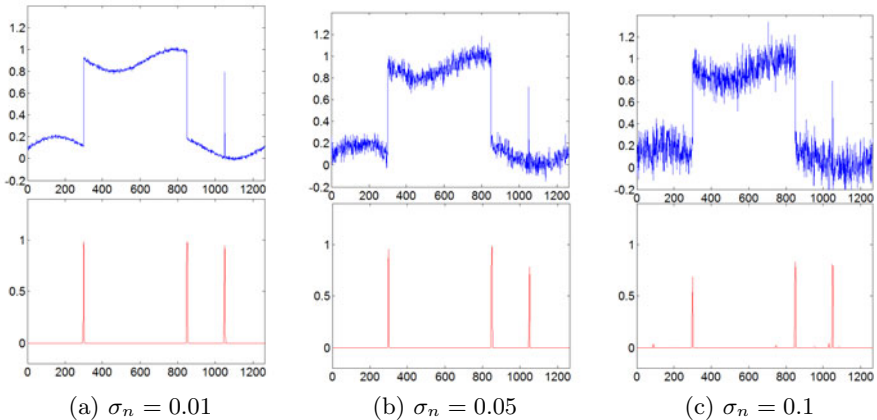


Fig. 6. LPC computation in the presence of additive white Gaussian noise, with noise standard deviation equaling (a) $\sigma_n = 0.01$, (b) $\sigma_n = 0.05$, and (c) $\sigma_n = 0.1$

The last experiment is concerned about the impact of noise on our LPC computation. Figure 6 shows the PC_2 function computed for a signal contaminated with

additive white Gaussian noise at three noise levels. It can be seen that the LPC computation successfully detects the distinctive features (edges and impulse) in all three cases, showing its strong robustness to noise (though the heights of LPC values may be moderately affected by heavy noise). This is another useful feature in practical applications, where many other techniques (e.g., derivative or gradient based edge detectors) are often sensitive to noise contaminations.

4 Conclusion

The purpose of this work is to extend the theory and methodology of local phase coherence, so that it can be converted to more practical techniques that can be applied to various signal processing applications for the analysis of signals and the detection of features. The major contribution of the current work as opposed to existing LPC computation is to formulate the problem using an optimization framework. Several technical issues have been studied in order to overcome a series of problems encountered in formulating the optimization problem and in finding the optimal solutions. The resulting LPC computation exhibits significantly broadened flexibilities such that it can be computed with arbitrary grouping of neighboring complex wavelet coefficients spread at any fractional scale ratios between successive scales. It also demonstrates strong robustness to noise. These flexibilities make our approach desirable in many potential applications, especially in the cases when multiple features exist and are close to each other, when only partial information of local phases is available, and/or when significant noise exists in the signal. Our future work is to apply the methodology developed in this work to practical signal and image applications, such as those in [12,13,14], so as to better exploit the advantages of LPC.

Acknowledgment

This research was supported in part by Natural Sciences and Engineering Research Council of Canada in the forms of Discovery, Strategic and CRD Grants, and by an Ontario Early Researcher Award, which are gratefully acknowledged.

References

1. Oppenheim, A.V., Lim, J.S.: The importance of phase in signals. *Proceedings of the IEEE* 69(5), 529–541 (1981)
2. Morrone, M.C., Burr, D.C.: Feature detection in human vision: a phase-dependent energy model. *Proceedings of the Royal Society of London, Series B* 235(128), 221–245 (1988)
3. Morrone, M.C., Owens, R.A.: Feature detection from local energy. *Pattern Recognition Letters* 6(5), 303–313 (1987)
4. Kovési, P.: Image features from phase congruency. *Journal of Computer Vision Research* 1(3), 1–26 (1999)

5. Fleet, D.J.: Phase-based disparity measurement. *CVGIP: Image Understanding* 53(2), 198–210 (1991)
6. Fleet, D.J., Jepson, A.D.: Computation of component image velocity from local phase information. *International Journal of Computer Vision* 5(1), 77–104 (1990)
7. Wang, Z., Li, Q.: Statistics of natural image sequences: temporal motion smoothness by local phase correlations. In: *Human Vision and Electronic Imaging XIV*, January 19-22. Proc. SPIE, vol. 7240 (2009)
8. Portilla, J., Simoncelli, E.P.: A Parametric Texture Model based on Joint Statistics of Complex Wavelet Coefficients. *International Journal of Computer Vision* 40, 49–71 (2000)
9. Daugman, J.: Statistical richness of visual phase information: update on recognizing persons by iris patterns. *International Journal of Computer Vision* 45(1), 25–38 (2001)
10. Zeng, K., Wang, Z.: Quality-aware video based on robust embedding of intra- and inter-frame reduced-reference features. In: *IEEE International Conference on Image Processing*, Hong Kong, China, September 26-29 (2010)
11. Wang, Z., Simoncelli, E.P.: Local phase coherence and the perception of blur. In: *Adv. Neural Information Processing Systems, NIPS 2003*, pp. 786–792. MIT Press, Cambridge (2004)
12. Hassen, R., Wang, Z., Salama, M.: Multi-sensor image registration based-on local phase coherence. In: *IEEE International Conference on Image Processing*, Cairo, Egypt, November 7-11 (2009)
13. Hassen, R., Wang, Z., Salama, M.: Multifocus image fusion using local phase coherence measurement. In: *International Conference on Image Analysis and Recognition*, Halifax, Canada, July 6-8 (2009)
14. Hassen, R., Wang, Z., Salama, M.: No-reference image sharpness assessment based on local phase coherence measurement. In: *IEEE International Conference on Acoustics, Speech, and Signal Processing*, Dallas, TX, March 14-19 (2010)

Edge Detection by Sliding Wedgelets

Agnieszka Lisowska

University of Silesia, Institute of Computer Science,
ul. Bedzinska 39, 41-200 Sosnowiec, Poland

alisow@ux2.math.us.edu.pl,

<http://www.math.us.edu.pl/al>

Abstract. In this paper the sliding wedgelet algorithm is presented together with its application to edge detection. The proposed method combines two theories: image filtering and geometrical edge detection. The algorithm works in the way that an image is filtered by a sliding window of different scales. Within the window the wedgelet is computed by the use of the fast moments-based method. Depending on the difference between two wedgelet parameters the edge is drawn. In effect, edges are detected geometrically and multiscale. The computational complexity of the sliding wedgelet algorithm is $O(N^2)$ for an image of size $N \times N$ pixels. The experiments confirmed the effectiveness of the proposed method, also in the application to noisy images.

Keywords: sliding wedgelets, edge detection, moments, multiresolution.

1 Introduction

Efficient edge detection is useful in many image processing tasks, like image segmentation, object recognition, etc. There is a wide spectrum of edge detection methods [1], [2], [3], [8], [12], [13], [14]. They can be classified into two groups — pointwise ones and geometrical ones. The methods from the first group are very fast and quite efficient. However, usually they are not noise resistant and cannot be used in some advanced image processing applications. So, in such issues like object recognition often the geometrical methods of edge detection are used. In such a case edges are represented by a set of line segments instead of a set of points. There are a few geometrical edge detection methods, the ones based on: the Radon transform [1], the wedgelet transform [13] or moments computation [12].

The use of geometrical multiscale methods in image processing has gained much attention recently. It follows from the fact that this approach reflects the way in which the human eye-brain system works. Indeed, the human eye can catch changes of location, scale and orientation [4], [7]. Many recently developed techniques of image processing are based on multiscale geometrical methods [5], [8], [9], [10], [11], [14], [15]. Especially, in the paper [13] the method of edge detection based on the wedgelet transform was proposed. However, the method is not fast, what excludes its use in real time applications.

In this paper the new multiscale geometrical edge detection method is proposed. The method inherits advantages of the pointwise and geometrical approaches to edge detection. It is based on image filtering on one hand and on a sliding wedgelet computation on the other hand. The notion of a sliding wedgelet, proposed in this paper, denotes that instead of computing wedgelets according to the image quadtree partition, like in the wedgelet transform, a wedgelet lying freely within the image is computed. The sliding wedgelet is used then in the definition of an image filter. So, because the algorithm is based on image filtering it is quite fast. Thanks to that it can be used in many advanced image processing or recognition methods.

2 Edge Detection Methods

There is a wide variety of edge detection methods. It is pointless to point out all of them. So, in this chapter the only methods which are helpful in the understanding of the proposed algorithm are presented. The pointwise methods are described in general, whereas the geometrical methods are represented by the wedgelet transform, since the proposed algorithm is based on it.

Image Filtering

Image filtering allows for image enhancement like edge detection, smoothing, etc. The filtered image is computed as a convolution of the original image $F : [0, 1] \times [0, 1] \rightarrow \mathbb{N}$ with a mask function $M : D \subset [0, 1] \times [0, 1] \rightarrow \mathbb{N}$. The typical size of the mask function is 3×3 or 5×5 pixels. The most commonly used filters are based on first or second derivatives like, for example, Sobel, Prewitt, Roberts or Canny filters [2], [3]. However, the more sophisticated results are obtained when the mask function M is nonlinear. In fact, the use of linear or nonlinear filter is determined by the application.

Wedgelet Transform

There are at least three geometrical methods of edge detection. The one based on the Radon transform [1], the second one based on moments computation [12] and the last one based on the wedgelet transform [13]. Since the latter one is multiscale it seems to be the most efficient one [13].

Consider an image $F : [0, 1] \times [0, 1] \rightarrow \mathbb{N}$. Consider then any square $D \subseteq [0, 1] \times [0, 1]$. Let us set the resolution of $[0, 1] \times [0, 1]$ as $N \times N$. It means that $[0, 1] \times [0, 1]$ can be represented by a matrix of $N \times N$ pixels of size $1/N \times 1/N$. It is not necessary to consider a squared domain but it simplifies the further considerations. Any straight line which connects two border pixels (not lying at the same side) is called a *beamlet* [8]. Then the characteristic function of the domain, bounded by the borders and the beamlet b , is given by the formula

$$W(x, y) = \mathbf{1}_{y \leq b(x)}, \quad (x, y) \in [0, 1] \times [0, 1] \quad (1)$$

and is called a *wedgelet* [6].

The definition of the wedgelet transform is based on a dictionary of wedgelets and an image quadtree partition. In more details, the dictionary of wedgelets is build of wedgelets of different positions, scales and orientations. For each tree node of the image quadtree partition the optimal wedgelet in Mean Square Error (MSE) sense is found. By applying the bottom-up tree pruning algorithm the wedgelet approximation is determined [6]. By using moments the computational complexity of the wedgelet transform can be $O(N^2 \log_2 N)$ for an image of size $N \times N$ pixels [16].

The edge detection method based on the wedgelet transform is defined as follows. First, the wedgelet image approximation is found. Second, the beamlets of that approximation are drawn instead of wedgelets. But, to avoid false edges, the only beamlets are drawn for which the difference between two wedgelet colors is larger than the fixed threshold. More details of the described method can be found in [13].

3 Sliding Wedgelets

Wedgelet filtering combines two theories — image filtering and the wedgelet transform. The aim of this paper is to release wedgelets from their fixed locations following from the wedgelet transform. In fact, the wedgelet transform is related to image quadtree partition. It causes that, for a fixed wedgelet's size, the only nonoverlapping correlations between a wedgelet and an image are determined. Additionally, the locations of appropriate wedgelets are fixed. From the edge detection point of view it is very inconvenient situation. In Fig. 1(a) the image with denoted two wedgelets of size 64×64 pixels from the wedgelet transform (so, their locations are fixed) representing the edge defined by the bird's wing is presented. As one can see the edge is not represented properly. In Fig. 1(b) the same image with denoted one sliding wedgelet of the same size is presented. One can easily see that the sliding wedgelet can represent the edge more efficiently than the ones from the wedgelet transform. It follows from the fact that its position may be chosen freely.

The algorithm of edge detection by sliding wedgelets is defined as follows.

Algorithm 1. Edge detection by sliding wedgelets

```

1. Input image F;
2. fix: size, shift, T;
3. for (i=0; i+size<=ImageSize; i+=shift)
4.     for (j=0; j+size<=ImageSize; j+=shift)
5.         compute wedgelet(F, i, j, size);
6.         if abs(c1-c2)>T
7.             draw beamlet(i, j, size);

```

The algorithm is quite simple in construction. However, some instructions of the presented code should be explained. Parameters *size*, *shift* and *T* denote a wedgelet size, a shift of sliding and a threshold for removing false edges, respectively. They are fixed by a user. The wedgelet in line 5 is computed with the

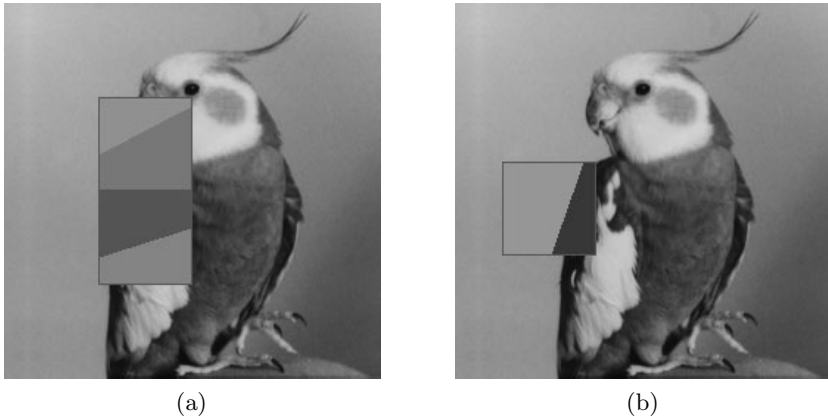


Fig. 1. (a) The edge represented by two wedgelets from the wedgelet transform, (b) the edge represented by one sliding wedgelet

help of the fast method proposed in [16], based on moments. Parameters c_1 , c_2 are the colors intensities of the wedgelet computed in line 5. The *if* statement in line 6 is necessary in order to avoid false edges drawing. In Fig. 2 the examples of detected edges for different sets of parameters are presented.

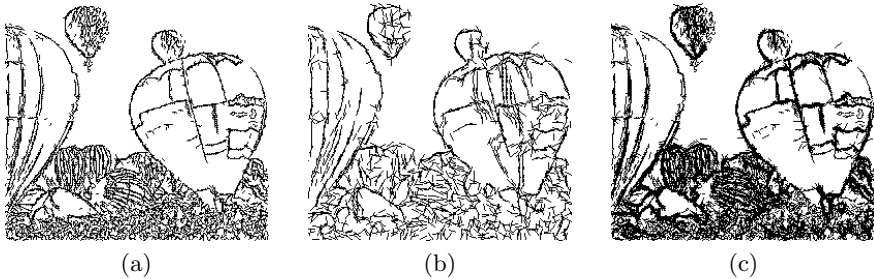


Fig. 2. The edges for a block of size: (a) 4×4 pixels, $shift = 2$, $T = 20$; (b) 8×8 pixels, $shift = 4$, $T = 20$; (c) 4×4 combined with 8×8 pixels, $shift = 2$ for both scales, $T = 20$ and $T = 40$, respectively

The computational complexity of the proposed method is linear. It follows from the use of a sliding window and the use of the fast wedgelet computation. From the sliding window construction follows that in the worst case ($shift=1$) the wedgelet parameters can be computed at most N^2 times for an image of size $N \times N$ pixels. From the definition of the fast wedgelet transform proposed in [16] follows that the wedgelet parameters (c_1 , c_2 and the beamlet end points) can be computed using $O(size^2)$ operations. So, the computational complexity of the proposed algorithm is $O(size^2 N^2)$. However, since $size$ is bounded by 32 in practical applications, the final computational complexity for the sliding wedgelet algorithm is $O(N^2)$.

In practice, the computations take less than 0.5 sec. for an image of size 256×256 pixels on a Pentium IV 3GHz processor.

4 Experimental Results

In this section some examples of edge detection by sliding wedgelets are presented. For comparison purposes also the results of Canny edge detector (from the Matlab toolbox) and the wedgelet transform-based edge detector (implemented by the author) are presented. The parameters in all three methods were chosen in the way to obtain the best possible results of edge detection. In Fig. 3 the standard benchmark images used in the experiments are presented.



Fig. 3. The standard set of benchmark images

In Fig. 4 the results of all three mentioned methods are presented. By analyzing the results one can conclude that the edges detected by the sliding wedgelets algorithm seems to be the most pleasant. The use of Canny edge detector leads to many false edges, omitting the fact that it is not any geometrical detector. In the case of edges detected with the help of the wedgelet transform one obtains usually not continuous edges. The edges produced from the proposed method are of different thickness. In some applications it can be good, since edges are usually of different thickness within an image. But in some other applications it can be cumbersome when one is looking for a single line representing an edge. Anyway, the proposed method detected the most of the details in comparison to the related methods.

In order to test the noise resistance of the proposed method, in Fig. 5 the results of edge detection of images contaminated by Gaussian noise with mean $M = 0$ and variance $V = 0.005$ are presented. The noise was added artificially



Fig. 4. Examples of edge detection: (a) Canny, (b) wedgelet transform, (c) sliding wedgelets

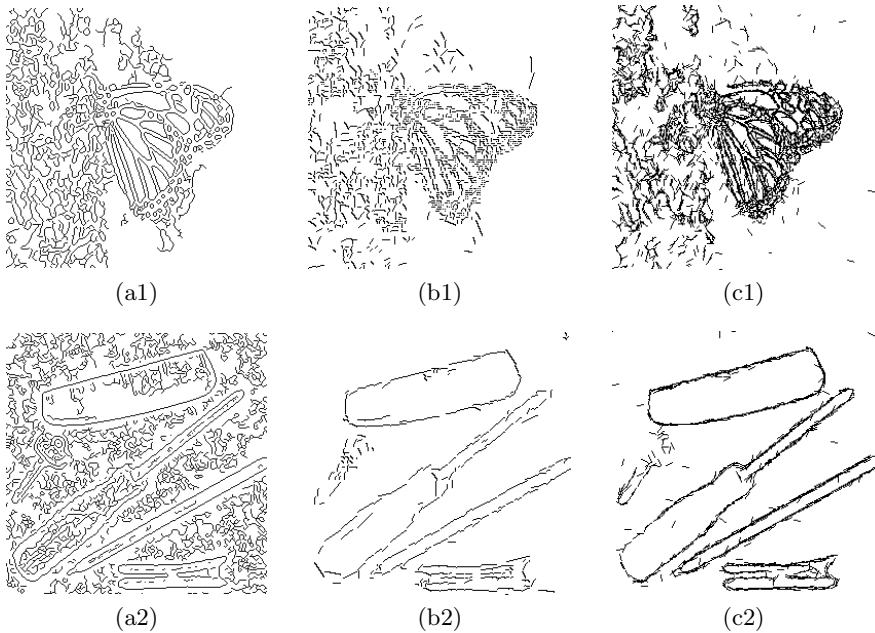


Fig. 5. Examples of edge detection from images contaminated by Gaussian noise with $M = 0$ and $V = 0.005$: (a) Canny, (b) wedgelet transform, (c) sliding wedgelets

with the use of the Matlab image processing toolbox. As one can expect the Canny detector is not noise resistant, so the results are not satisfactory. The proposed method copes quite well with the noise. However, some false edges appeared in the result images. The wedgelet transform-based method seems to give slightly lesser number of false edges. But, on the other hand, the real edges are more pleasant visually in the sliding wedgelet algorithm.

5 Summary

In the paper the new method of image filtering has been presented with the application to edge detection. The method combines two different theories of edge detection — image filtering and geometrical edge detection. From all that follows that the proposed method is fast and detects edges in multiscale geometrical way.

It is important to note that one of the advantages of the proposed algorithm, the low computational complexity, follows from the theory of moments. By applying moments theory to the wedgelet transform computation one can significantly reduce its computational complexity [16]. Without this approach the algorithm proposed in this paper could not be fast.

References

1. Deans, S.R.: The Radon Transform and Some of Its Applications. John Wiley and Sons, New York (1983)
2. Canny, J.: Computational Approach To Edge Detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8, 679–714 (1986)
3. Gonzalez, R., Woods, R.: *Digital Image Processing*. Addison-Wesley, Reading (1992)
4. Olshausen, B.A., Field, D.J.: Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. *Nature* 381, 607–609 (1996)
5. Meyer, F.G., Coifman, R.R.: Brushlets: A Tool for Directional Image Analysis and Image Compression. *Applied and Computational Harmonic Analysis* 4, 147–187 (1997)
6. Donoho, D.L.: Wedgelets: Nearly-minimax estimation of edges. *Annals of Statistics* 27, 859–897 (1999)
7. Humphreys, G.W. (ed.): *Case Studies in the Neuropsychology of Vision*. Psychology Press, UK (1999)
8. Donoho, D.L., Huo, X.: Beamlet Pyramids: A New Form of Multiresolution Analysis, Suited for Extracting Lines, Curves and Objects from Very Noisy Image Data. In: *Proceedings of SPIE*, vol. 4119 (2000)
9. Demaret, L., Friedrich, F., Führ, H., Szygowski, T.: Multiscale Wedgelet Denoising Algorithms. In: *Proceedings of SPIE*, vol. 5914, pp. 1–12 (2005)
10. Labate, D., Lim, W., Kutyniok, G., Weiss, G.: Sparse Multidimensional Representation Using Shearlets. In: *Proceedings of the SPIE*, vol. 5914, pp. 254–262 (2005)
11. Mallat, S., Pennec, E.: Sparse Geometric Image Representation with Bandelets. *IEEE Transactions on Image Processing* 14(4), 423–438 (2005)
12. Popovici, I., Withers, W.D.: Custom-Built Moments for Edge Location. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(4), 637–642 (2006)
13. Lisowska, A.: Geometrical Multiscale Noise Resistant Method of Edge Detection. In: Campilho, A., Kamel, M.S. (eds.) *ICIAR 2008*. LNCS, vol. 5112, pp. 182–191. Springer, Heidelberg (2008)
14. Lisowska, A.: Multiscale Moments-Based Edge Detection. In: *Proceedings of EUROCON 2009 Conference*, St.Petersburg, Russia, pp. 1414–1419 (2009)
15. Mallat, S.: Geometrical Grouplets. *Applied and Computational Harmonic Analysis* 26(2), 161–180 (2009)
16. Lisowska, A.: Moments-Based Fast Wedgelet Transform. *Journal on Mathematical Imaging and Vision* 39(2), 180–192 (2011)

Adaptive Non-linear Diffusion in Wavelet Domain

Ajay K. Mandava and Emma E. Regentova

Electrical and Computer Engineering
University of Nevada, Las Vegas
4505 Maryland Parkway, Box 454026
Las Vegas, NV 89154-4026
mandavaa@unlv.nevada.edu, emma.regentova@unlv.edu

Abstract. Traditional diffusivity based denoising models detect edges by the gradients of intensities, and thus are easily affected by noise. In this paper, we develop a nonlinear diffusion denoising method which adapts to the local context and thus preserves edges and diffuses more in the smooth regions. In the proposed diffusion model, the modulus of gradient in a diffusivity function is substituted by the modulus of a wavelet detail coefficient and the diffusion of wavelet coefficients is performed based on the local context. The local context is derived directly by analyzing the energy of transform across the scales and thus it performs efficiently in the real-time. The redundant representation of the stationary wavelet transform (SWT) and its shift-invariance lend themselves to edge detection and denoising applications. The proposed stationary wavelet context-based diffusivity (SWCD) model produces a better quality image compared to that attained by two high performance diffusion models, i.e. higher Peak Signal-to-Noise Ratio on average and lesser artifacts and blur are observed in a number of images representing texture, strong edges and smooth backgrounds.

Keywords: Stationary Wavelet Transform; Non-linear Diffusion; Context-based Denoising.

1 Introduction

The need for efficient image restoration methods has grown with the massive production of digital images and acquisition systems of all kinds. Among denoising methods the non-linear diffusion represents a simple yet efficient approach. The basic idea behind nonlinear diffusion filtering is to obtain a family $u(x, t)$ of filtered versions of the signal $f(x)$ as a solution of a suitable diffusion process

$$u_t = (g(|u_x|) u_x)_x \quad (1)$$

with $f(\cdot)$ as an initial condition:

$$u(x, 0) = f(x) \quad (2)$$

Here subscripts denote partial derivatives, and the diffusion time t is a simplification parameter with larger values corresponding to stronger filtering. The

diffusivity $g(|u_x|)$ is a nonnegative function that controls the amount of diffusion. Usually it is decreasing in $|u_x|$. This ensures that strong edges are less blurred by the diffusion filter than the noise and low-contrast details. Depending on the choice of the diffusivity function, equation (1) covers a variety of filters. Some of the nonlinear anisotropic diffusion techniques are Perona–Malik filter [1], Weickert filter [2,3], Vogel-Omans's [4] and Rudin-Osher-Fatemi's [5] total variation diffusion. These techniques rely on the diffusion flux to iteratively eliminate small variations due to noise or texture, and to preserve large variations due to edges. For the multiplicative noisy image, however, the general signal/noise relationship no longer exists, since the variations due to noise may be larger than those due to signal. This limits the application of nonlinear diffusion methods for image processing. Nonlinear diffusion techniques rely on the gradient operator to distinguish signal from noise. Such a method often cannot achieve a precise separation of signal and noise. Image denoising problems are better solved if a powerful signal/noise separating tool such as for example, wavelet analysis is incorporated in the diffusion process.

Recent work [6-13] has shown that nonlinear anisotropic diffusion can be employed within the framework of the dyadic wavelet transform (DWT). We refer to the integration of nonlinear diffusion and wavelet shrinkage as wavelet diffusion. This approach has more favorable denoising properties than nonlinear diffusion in the intensity domain. It is also distinguished from wavelet-based denoising methods such as wavelet shrinkage by its improved edge-enhancement and iterative noise reduction features.

In [6,] a nonlinear multiscale wavelet diffusion method for the ultrasound speckle suppression and edge enhancement is presented. The edges are detected using normalized wavelet modulus and speckle is suppressed by employing the iterative multiscale diffusion of wavelet coefficients. The diffusion threshold is estimated from the normalized modulus in the homogenous speckle regions, in order to adapt to the noise variation with iteration. The automatic identification of homogenous regions is implemented using two-stage classification. Although the method could reduce the speckle and preserve edges, the low-contrast edges are blurred significantly.

In [7], Bruni proposed another wavelet and partial differential equation (PDE) model for image denoising. Wavelet coefficients are modeled as waves that grow while expanding along scales. The model establishes a precise link between corresponding modulus maxima in the wavelet domain and then allows predicting wavelet coefficients at each scale from the first one from waves obeying a precise partial differential equation. This property combined with theoretical results about the characterization of singularities in the wavelet domain enables to discard noise. A drawback of this model are artifacts and the computational cost.

Shih and Liao [8] addressed a single step nonlinear diffusion that can be considered equivalent to a single shrinkage iteration of coefficients of Mallat's Zhong dyadic wavelet transform (MZ-DWT) [9]. Nonlinear diffusion begins with a gradient operator, which may be badly influenced by the noise present in the image. The characteristics of wavelet transform to obtain an edge estimate makes it less sensitive to noise i.e. to correctly separate the high frequency components from the low frequency ones and to retain the values of the high-frequency components that

corresponds to those having larger magnitudes and on the other hand to suppress those having smaller magnitudes. However, the method does not consider a context information and as a result is not free from artifacts.

Bao and Krim [10] addressed the problem of texture losses in diffusion process in scale spaces by incorporating ideas from wavelet analysis. They showed that using wavelet frames of higher order than Haar's is as good as to accounting for longer term correlation structure, while preserving the local focus on equally important features and illustrated the advantages of removing noise while preserving features.

Mrazek and Weickert [11] have analyzed correspondences between explicit one-dimensional schemes for nonlinear diffusion and discrete translation-invariant Haar wavelet shrinkage. Weickert et al. [11, 12] describe connections between discrete diffusion filtering and Haar wavelet shrinkage, including a locally analytic four-pixel scheme, but focused on the 1-D or the isotropic 2-D case with a scalar-valued diffusivity. This method enhances the edge but doesn't preserve the object shape.

In [13], Chen developed three denoising schemes by combining PDE with wavelets. In the first proposed model, the diffusion is a function of the Rudin-Osher-Fatemi's total variation model and used amount of advection to diffuse differently in various directions and the largest amount of advection occurs in the normal direction and the smallest in the tangent direction. The model could preserve edges better and displayed strong noise resistance.

In the above discussed methods no information about the local context is taken into account and no differentiation is made between texture and extended objects edges. In this paper, we combine in a single framework the advantages of non-linear diffusion and multiresolution decomposition and explore the context information to control the diffusion. The diffusivity function is used as a weighting function to the wavelet coefficients of a stationary wavelet transform (SWT) which provides both scale invariance and context information. The latter is derived from the transform energy observed locally spatially and across the scales. We compare the performance of the proposed method to the Weickert's diffusivity and the method in [8]. Section 2 provides a theoretical background and introduces the new local context based diffusion in the stationary wavelet domain (SWCD). Section 3 shows the experimental results.

2 Local Context Based Diffusion in Stationary Wavelet Domain (SWCD)

In a decimated discrete wavelet transform (DWT) after high and low pass filtering, coefficients are down sampled. Although this prevents redundancy and allows for using a same pair of filters in different levels, this decimated transform lacks shift invariance. Thus, small shifts in the input signal can cause major variations in the distribution of energy of coefficients at deferent levels. Even with periodic signal extension, the DWT of a translated version of a signal X is not, in general, the translated version of the DWT of X . To restore the translation invariance one can average a slightly different DWT, called ε -decimated DWT, to define the stationary wavelet transform (SWT) [14]. SWT algorithm is simple; the decimated DWT for a

given signal can be obtained by convolving the signal with the appropriate filters as in the DWT case but without down-sampling. The two-dimensional SWT leads to a decomposition of approximation coefficients at level j to four components: the approximation at level $j+1$, and the details in three orientations, i.e., horizontal, vertical, and diagonal). Considering the multi-sampling filter banks, SWT decomposition is shown in Eq.3.

$$\begin{aligned}
A_{j,k_1,k_2} &= \sum_{n_1} \sum_{n_2} h_0^{\uparrow 2^j}(n_1 - 2k_1) h_0^{\uparrow 2^j}(n_2 - 2k_2) A_{j-1,n_1,n_2} \\
D_{j,k_1,k_2}^1 &= \sum_{n_1} \sum_{n_2} h_0^{\uparrow 2^j}(n_1 - 2k_1) h_1^{\uparrow 2^j}(n_2 - 2k_2) A_{j-1,n_1,n_2} \\
D_{j,k_1,k_2}^2 &= \sum_{n_1} \sum_{n_2} h_1^{\uparrow 2^j}(n_1 - 2k_1) h_0^{\uparrow 2^j}(n_2 - 2k_2) A_{j-1,n_1,n_2} \\
D_{j,k_1,k_2}^3 &= \sum_{n_1} \sum_{n_2} h_1^{\uparrow 2^j}(n_1 - 2k_1) h_1^{\uparrow 2^j}(n_2 - 2k_2) A_{j-1,n_1,n_2} \quad (3)
\end{aligned}$$

Where $h_0^{\uparrow 2^j}, h_1^{\uparrow 2^j}$ denote the $(2^j - 1)$ zeros padded between h_0 and h_1 , respectively. The inverse transform of SWT follows Eq.4.

$$\begin{aligned}
A_{j-1,n_1,n_2} &= \frac{1}{4} \sum_{i=0}^3 \{ \sum_{k_1} \sum_{k_2} g_0(n_1 - 2k_1 - i) g_0(n_2 - 2k_2 - i) A_{j,k_1,k_2} \\
&\quad + \sum_{k_1} \sum_{k_2} g_0(n_1 - 2k_1 - i) g_1(n_2 - 2k_2 - i) D_{j,k_1,k_2}^1 \\
&\quad + \sum_{k_1} \sum_{k_2} g_1(n_1 - 2k_1 - i) g_0(n_2 - 2k_2 - i) D_{j,k_1,k_2}^2 \\
&\quad + \sum_{k_1} \sum_{k_2} g_1(n_1 - 2k_1 - i) g_1(n_2 - 2k_2 - i) D_{j,k_1,k_2}^3 \} \quad (4)
\end{aligned}$$

where A and D are approximation and detail coefficients, respectively.

From the above two equations, we can verify that SWT includes redundant information and shift-invariance suitable for structure analyses and denoising. Smooth regions in image are represented mainly by approximation coefficients. Level 1 and Level 2 detail subbands convey noise and the fine-grain texture information. The higher scales carry the information of edges of extended objects. To perform diffusion selectively and adaptively the local structure is to be observed from the distribution of energy of transform as it carries important information about the local context. Consider two-level Haar SWT of the noisy image with Haar wavelet. The energy of transform in respective subbands is defined as follows:

$$\begin{aligned}
E_{S,nxm}^V &= \sum_{i=1}^n \sum_{j=1}^m |D_{i,j}^{(V)}|^2, \quad E_{S,mxn}^D = \sum_{i=1}^n \sum_{j=1}^m |D_{i,j}^{(D)}|^2 \quad \text{and} \\
E_{S,mxn}^H &= \sum_{i=1}^n \sum_{j=1}^m |D_{i,j}^{(H)}|^2 \quad (5)
\end{aligned}$$

where $m \times n$ is a window at scale s , and k indicates the subband, i.e., V- vertical, D- diagonal and H-horizontal.

Fig.1 shows examples per context. The ratio in the figure is calculated in sliding windows of size of 9×9 pixels based on cumulative energies in three subbands; it is normalized and plotted versus different noise levels, i.e. $\sigma = 10, 20, 30, 40$ of Gaussian white noise. This ratio characterizes the local context for controlling the diffusion equation. Specifically, smooth regions affected by noise can be identified and let undergoing larger diffusion; edges of texture and extended objects exhibit ratio values different from that of smooth regions and thus can be lesser/slower diffused.

Thus, the ratio of energies can be applied as an additional factor controlling the diffusion.

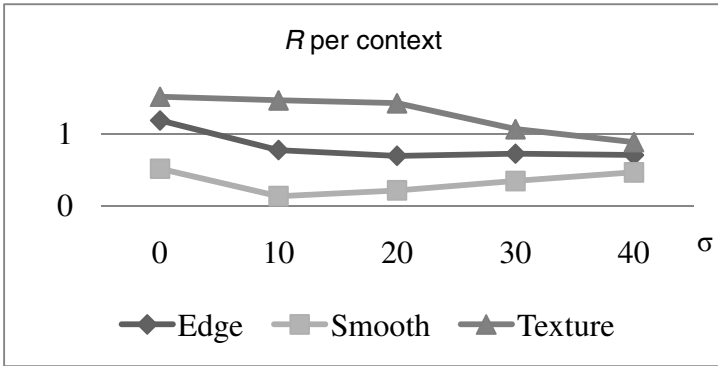


Fig. 1. Distribution of E_2/E_1 for different contexts vs Gaussian white noise $\sigma= 10, 20, 30, 40$

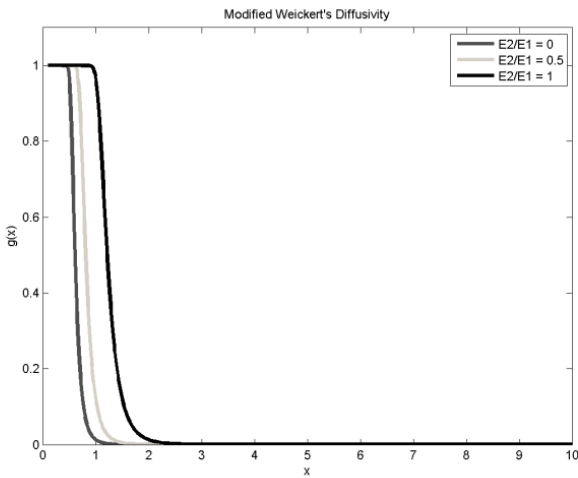


Fig. 2. Stability graph for modified Weickert's diffusivity function

As in Shih's method [8], we perform diffusion at level 1 and make related to the ratio of $E2/E1$. In Weickert's diffusivity function we introduce a coefficient R which is a ratio of energies in the wavelet subbands. This forms a modified diffusivity function as in Eq 6.

$$g(|s|) = 1 - \exp\left(\frac{-3.315}{\left(\frac{|s|}{\lambda*(2-R)}\right)^8}\right) \quad (6)$$

Where $R = E2/E1$, and if $E1 = 0$, then it is replaced with $\epsilon = 0.001$.

Fig. 2 illustrates the bounded diffusivity for above modified Weickert's diffusivity equation.

In the diffusivity function, $|s(x,y)|$ is the edge estimate at pixel (x,y) , given by $|s_i(x,y)| = |D_i^{(k)}(x,y)|$, where $i = \{1,2,3\}$ and $k = \{v,h,d\}$. Diffusion is performed as $D_i^{(k)} = cD_i^{(k)*p(|s_i|)$ i. e. $g(|s_i|) = 1-p(|s_i|)$ and the image is reconstructed.

3 Experiment

To study the performance of SWCD we choose Gaussian white noise with standard deviation $\sigma = 10, 20, 30, 40$. The evaluation is performed based on PSNR according to Eq.7, where MSE- is a mean square error:

$$\text{PSNR} = 10 \log \frac{255^2}{\text{MSE}} \quad (7)$$

For comparison we select Weickert's and Shih's [8] diffusivities with the latter using Weickert's diffusivity function applied directly to the first level of horizontal and vertical subbands of DWT with Haar wavelet function.

In Fig.3 we show the original Lena image with an area chosen in a relatively smooth part of the image and enlarged marked area with the Gaussian white noise of $\sigma = 10$, and results of diffusion for various iterations (iter = 5, 10 and 15) and $\lambda = 10$. We notice that the method as in [8] produces artifacts which are not seen in the results of the proposed approach. In Fig.4 we show the original texture image and results of diffusion in a certain area. We notice that the Weickert's diffusivity and method as in [8] produce artifacts and blur edges in a greater extent but the proposed approach preserves edges of both texture regions and extended objects. In Cameraman image on Fig. 5 we see that the cameraman's face and camera's small details are vivid after diffusion by the proposed technique, and the background is smoother. In Fig. 6 that is in Pepper image, smooth pepper surface and the pepper contour are better preserved by the proposed technique. Fig.7 show PSNR for different noise levels (Gaussian, $\sigma = 10, 20, 30, 40$) for six test images, Weickert's diffusivity with $\lambda = 10$ and 10 iterations, method as in [8] with $\lambda = 10$ and 10 iterations, and the herein proposed SWCD method with $\lambda = 100$ and single iteration. From the results it can be observed that SWCD performs better in terms of both subjective quality and objective measures compared to other two counterpart techniques.

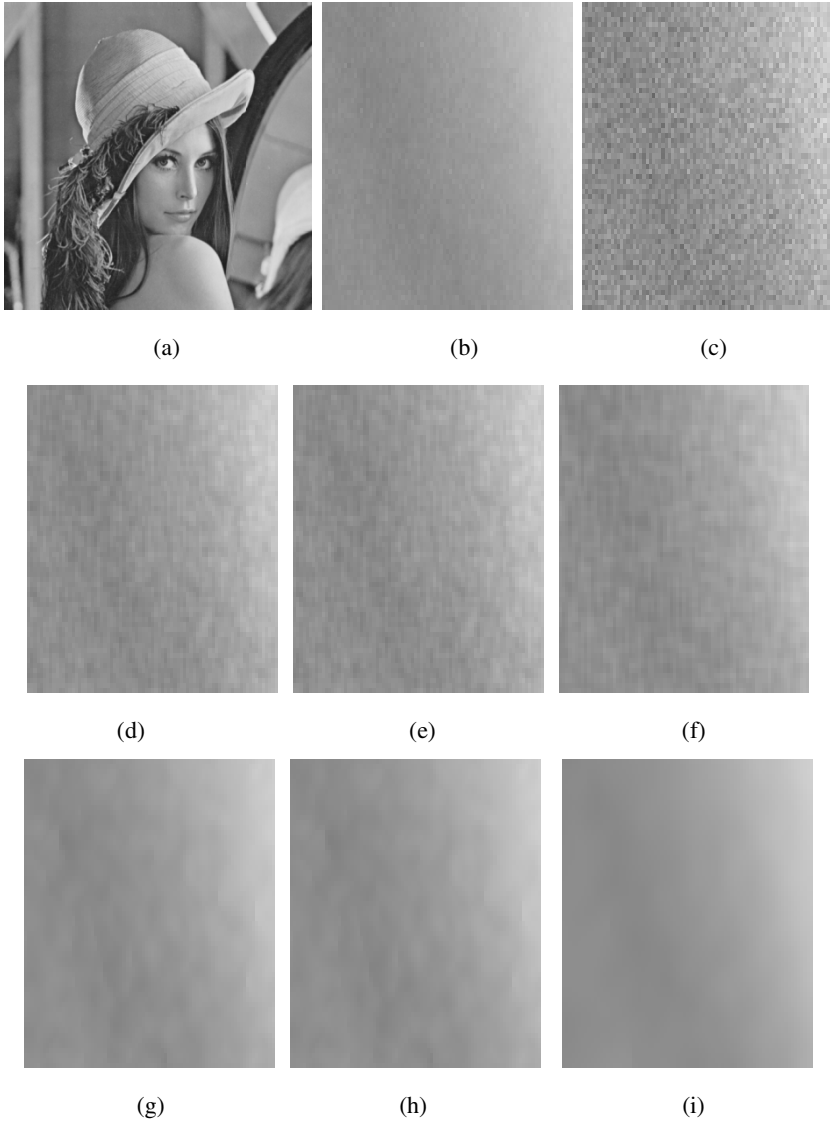


Fig. 3. a) Original Image b) Part of the original c) Image with noise (PSNR = 28.12 dB) d) Method as in [8] (PSNR = 29.32 dB & iter = 5) e) Method as in [8] (PSNR = 28.01 dB & iter = 10) f) Method as in [8] (PSNR = 27.35 dB & iter = 15) g) SWCD (PSNR = 32.47 dB & iter = 5) h) SWCD (PSNR = 31.08 dB & iter = 10) i) SWCD (PSNR = 30.11 dB & iter = 15)

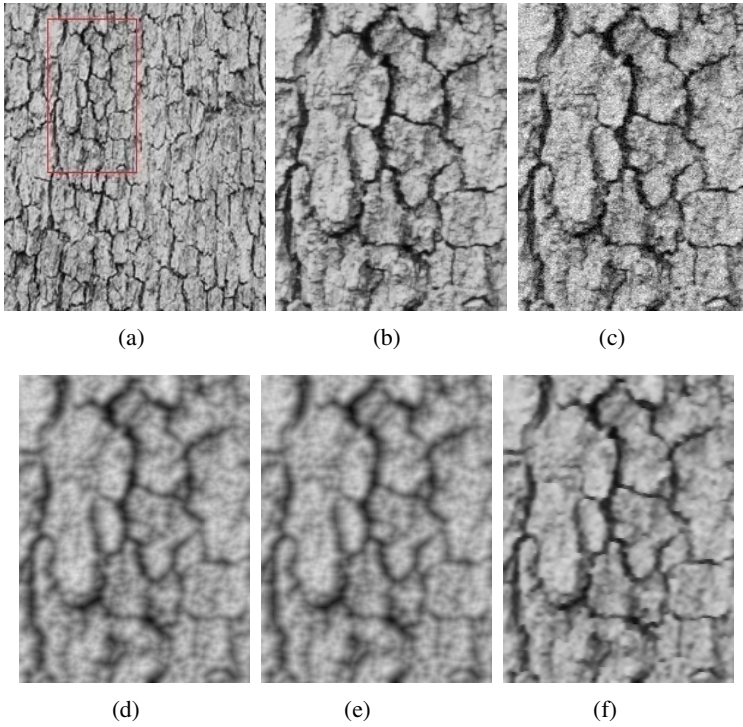


Fig. 4. a) Original Image b) Part of the original c) Image with noise (PSNR = 22.11 dB) d) Weickert's Diffusivity model (PSNR = 21.19 dB) e) Method as in [8] (PSNR = 20.51 dB) f) SWCD (PSNR = 23.56 dB)

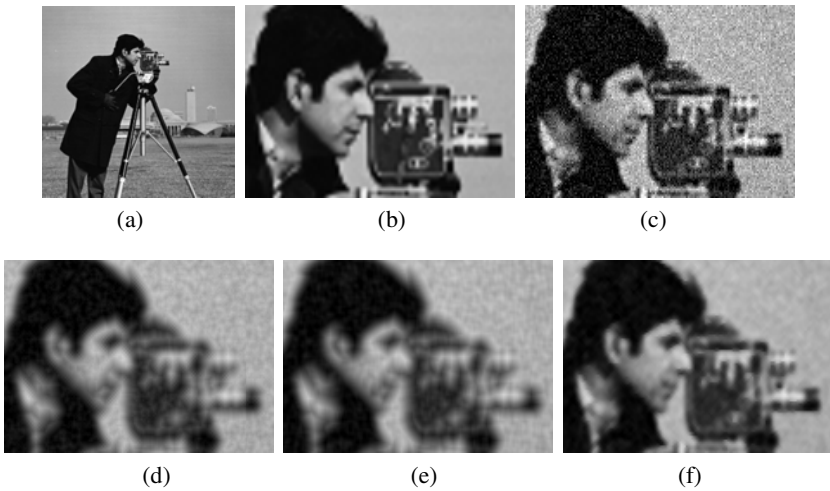


Fig. 5. a) Original b) Part of the original c) Image with noise (PSNR = 22.12 dB) d) Weickert's Diffusivity (PSNR = 26.51 dB) e) Method as in [8] (PSNR = 25.95 dB) f) SWCD (PSNR = 29.76 dB)

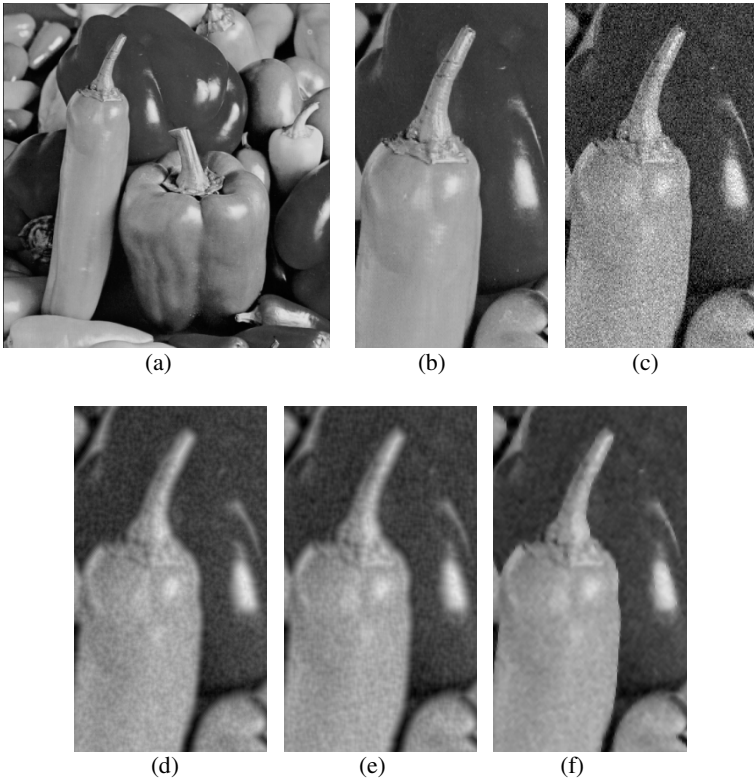


Fig.6. a) Original Image b) Part of the original c) Image with noise (PSNR = 22.13 dB) d) Weickert's Diffusivity (PSNR = 26.95 dB) e) Method as in [8] (PSNR = 26.40 dB) f) SWCD (PSNR = 30.03 dB)

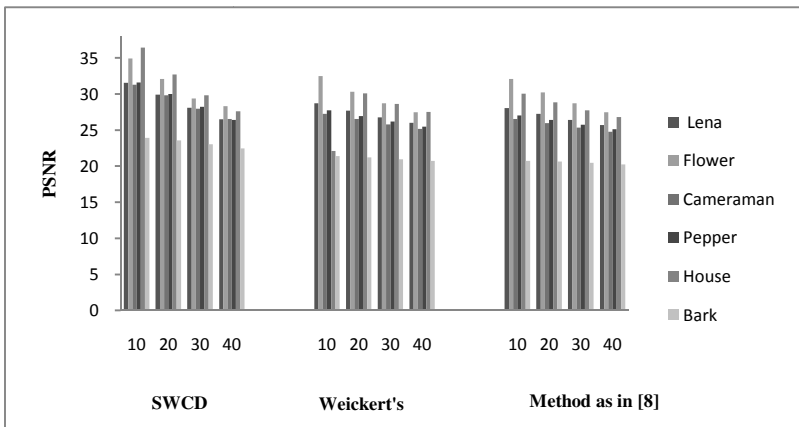


Fig.7. PSNR for denoising of Gaussian white noise ($\sigma = 10, 20, 30, 40$)

4 Conclusion

The paper presented an adaptive image denoising method based on non-linear diffusion in the wavelet domain. The wavelet transform is stationary, i.e. redundant shift invariant shown to be effective for denoising. The magnitude of diffusion is controlled adaptively by the local context measured by the ratio of transform energies at scales 2 and 1. Unlike other context-based denoising models, here neither segmentation nor edge detection is performed prior to denoising; and thus method can be implemented in the real time. Based on the evaluation results, the SWCD shows a higher on average PSNR and perceptual quality compared to those of two reference methods.

Acknowledgments. This work is partially supported by Faculty Sabbatical Leave program of University of Nevada, Las Vegas.

References

1. Perona, P., Malik, J.: Scale-space and edge detection using anisotropic diffusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 12, 629–639 (1990), doi:10.1109/34.56205
2. Weickert, J., Steidl, G., Mrazek, P., Welk, M., Brox, T.: Diffusion filters and wavelets: What can they learn from each other? In: Paragios, N., Chen, Y., Faugeras, O. (eds.) *The Handbook of Mathematical Models in Computer Vision*. Springer, New York (2005)
3. Weickert, J.: *Anisotropic Diffusion in image processing*. ECMI Series. Teubner, Stuttgart (1998)
4. Vogel, C., Oman, M.: Fast Robust Total Variation Based Reconstruction of Noisy Blurred Images. *IEEE Transactions on Image Processing* 7, 813–824 (1998), doi:10.1109/83.679423
5. Rudin, L., Osher, S., Fatemi, E.: Nonlinear total variation based noise removal algorithms. *Physica D* 60, 259–268 (1992), doi:10.1016/0167-2789(92)90242-F
6. Yue, Y., Croitoru, M.M., Bidani, A., Zwischenberger, J.B., Clark Jr., J.W.: Nonlinear Multiscale Wavelet Diffusion for Speckle Suppression and Edge Enhancement in Ultrasound Images. *IEEE Transactions on Medical Imaging* 25, 297–311 (2006), doi:10.1109/TMI.2005.862737
7. Bruni, V., Piccoliand, B., Vitulano, D.: Wavelets and partial differential equations for image denoising. *Electronic Letters on Computer Vision and Image Analysis* 6, 36–53 (2008)
8. Shih, A.C.-C., Liao, H.-Y.M., Lu, C.-S.: A New Iterated Two-Band Diffusion Equation: Theory and Its Applications. *IEEE Transactions on Image Processing* (2003), doi: 10.1109/TIP.2003.809017
9. Mallat, S., Zhong, S.: Characterization of Signals from Multiscale Edges. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 14, 710–732 (1992), doi:10.1109/34.142909
10. Bao, Y., Krim, H.: Towards bridging scale-space and multiscale frame analyses. In: Petrosian, A.A., Meyer, F.G. (eds.) *Wavelets in Signal and Image Analysis*. Computational Imaging and Vision, vol. 19, ch. 6. Kluwer, Dordrecht (2001)

11. Mrazek, P., Weickert, J., Steidl, G.: Diffusion inspired shrinkage functions and stability results for wavelet denoising. *Int. J. Computer Vision* 64, 171–186 (2005), doi:10.1007/s11263-005-1842-y
12. Welk, M., Weickert, J., Steidl, G.: A four-pixel scheme for singular differential equations. In: Kimmel, R., Sochen, N. (eds.) *Scale-Space 2005*. LNCS, vol. 3459, pp. 585–597. Springer, Heidelberg (2005), doi:10.1007/11408031_52
13. Chen, L.: Image De-noising Algorithms Based on PDE and Wavelet, *iscid*. In: 2008 International Symposium on Computational Intelligence and Design, vol. 1, pp. 549–552 (2008), doi:10.1109/ISCID.2008.196
14. Nason, G.P., Silverman, B.W.: The stationary wavelet transform and some statistical applications. *Lecture Notes in Statistics*, vol. 103, pp. 281–299 (1995)

Wavelet Domain Blur Invariants for 1D Discrete Signals

Iman Makaremi, Karl Leboeuf, and Majid Ahmadi

Department of Electrical and Computer Engineering, University of Windsor,
401 Sunset Ave, Windsor, ON, N9B 3P4, Canada
{makarem, leboeu3, ahmadi}@uwindsor.ca

Abstract. Wavelet domain blur invariants, which were proposed for the first time in [10] by the authors, are modified in order to suit a wider range of applications. With the modified blur invariants, it is possible to address the applications in which the blur systems are not necessarily energy-preserving. Also, for a simpler implementation of the wavelet decomposition for discrete signals, we use a method which preserves an important property of the invariants: shift invariance. The modified invariants are utilized in two different experiments in order to evaluate their performance.

Keywords: Blur Moment Invariants, Direct Analysis, Feature Extraction, Shift Invariant Wavelet Transform.

1 Introduction

Perfection is nearly impossible when it comes to signal acquisition. Different sources of degradation cause the acquired signal to not be exactly identical to the original one. The effect of some of these degradation sources is considerably high, which can vastly affect expected outcomes.

Blur is one of the degradations that could effectively reduce the discrimination power. It could be introduced to signals due to the movement of the subject or the data acquisition instrument. Also, the environment that the signal travels in could blur the signal to some extent. Blur can be modeled as a linear shift invariant system

$$y[n] = b * x[n] \quad (1)$$

where x and y are the original and blurred signals, respectively, and b is the blur system.

In most of the cases, the blur system is unknown or only partial information is available. The approaches in the literature that deal with removing the blur effect are mainly two different types: blind restoration and direct analysis. In blind restoration, the main purpose is to estimate the blur system and the original signal with partial information about the acquisition system. There are numerous proposed methods in literature for this type of approaches [8]. However, the main problems with them are that they usually require high computational effort, and the problem is usually ill-posed.

The direct analysis approaches, on the other hand, try to find characteristics of the original signal without going through estimating the blur system. Flusser et. al. [5][6] proposed the first direct analysis method based on the geometric moments. In order to simplify the problem, they made two assumptions: the blur system is centrally symmetric and energy-preserving. These assumptions are generally used in developing any other kind of blur invariant descriptors as well. Subsequently, they modified their method in order to make it invariant to translation, scaling, and change of contrast as well [4], and generalized it for N -dimensional signals [3]. Zhang et al. [15] employed Legendre moments for extraction of blur invariant descriptors in the spatial domain.

Along with the blur invariant descriptors proposed in the spatial domain, there are some other methods that are developed in the Fourier domain. The first invariants in this domain were proposed by Flusser and Suk [5]. They showed that the tangent of the Fourier transform phase is blur invariant. Ojansivu and Heikkil [13] also proposed their blur invariant features in this domain based on phase-only bispectrum.

Wavelet domain invariants were first proposed by the authors [10] with application in analyzing EEG, ECG, speech signals, and signals acquired in single point ultrasound measurements. Defining blur invariants in the wavelet domain provides the advantage of analyzing signals at different scales with different bases. It has been shown that Flusser's spatial domain blur invariants [5] are a special case of the wavelet domain blur invariants, and their limitation in spatial domain is not an issue in the wavelet domain [10]. Also in a comprehensive experiment, it was shown that the wavelet domain invariants are performing better than the conventional blur invariants.

In this paper, those invariants are made available for discrete wavelet decomposition, while preserving the shift invariance. Also, the invariants are modified in order to remove a restriction on the blur system: the energy-preserving property. In the next section, some of the basic definitions are reviewed and complimentary ones are proposed. In section 3, the discrete wavelet transform is reviewed, its limitation is discussed, and alternative approaches are explained. Section 4 is devoted to moments in the wavelet domain and how blur represents itself in this domain. Also, the modified blur invariants are proposed in this section. In section 5, the performance of the invariants are evaluated through two experiments. The paper is concluded in section 6.

2 Basic Definitions and Notation

In this section, some basic terms are defined and explained.

Definition 1. *The p^{th} order ordinary geometric moment of discrete signal x in the spatial domain is defined by*

$$m_p^x = \sum_n n^p x[n]. \quad (2)$$

Definition 2. The centroid of signal x is

$$c^x = \frac{m_1^x}{m_0^x} \quad (3)$$

Definition 3. The p^{th} order central moment of discrete signal x in the spatial domain is defined by

$$\mu_p^x = \sum_n (n - c^x)^p x[n]. \quad (4)$$

For (3) to hold, and for definition 3 to be valid, x is required to have a nonzero m_0^x . In the case that the moments of x are zero up to a certain order, the following definitions are proposed.

Definition 4. If the moments of signal x are zero up to order $M-1$, its centroid is defined as

$$\varsigma^x = \frac{m_{M+1}^x}{(M+1)m_M^x} \quad (5)$$

Definition 5. If the moments of discrete signal x are zero up to order $M-1$, its p^{th} order central moment ($p \geq M$) in the spatial domain is defined by

$$\mu_p^x = \sum_{n \in N} (n - \varsigma^x)^p x[n]. \quad (6)$$

Definitions 2 and 3 are special cases of definitions 4 and 5 respectively in which $M = 0$.

3 Shift Invariant Discrete Wavelet Transform

The use of wavelet transform as a powerful signal processing tool has several outstanding advantages over other signal processing techniques. Employing wavelet transform yields the opportunity to analyze signals at different times and frequencies simultaneously. There are also a large number of bases available for performing the transform which provides access to information that may not be extractable by other techniques.

Discrete wavelet transform (DWT) is introduced for analyzing discrete signals. However, it suffers from a major drawback: it is not shift invariant, and this is due to the dyadic sub-sampling [7]. In order to make the moments invariant to shift, it is necessary to have a shift invariant wavelet transform.

There have been several different techniques developed to produce shift invariant wavelet transforms. Continuous Wavelet Transform (CWT) does not suffer from the same drawback as its counterpart in the discrete domain [12]. Therefore, it is typically utilized when the wavelet transform is only required at a few specific scales. Mallat proposed a scheme [11] that is an approximation of CWT, which was later proved that is invariant to shift [14]. *À trous* algorithm [12] is the simplest and yet an effective technique that is proposed to make DWT

invariant to shift. In this technique, the sub-sampling operator is removed, and the filters are instead up-sampled at each level by inserting zeros between every two coefficients [12]. Shift invariance has been also achieved by calculating the wavelet transform of all shifts [9].

In this paper, we use the *à trous* algorithm which complies with our expectations the best. In this algorithm, scaling and wavelet filters at scale $j + 1$ are defined as

$$h_{j+1}[k] = h_j[k] \uparrow 2 = \begin{cases} h_j[\frac{k}{2}], & k \text{ even} \\ 0, & k \text{ odd} \end{cases} \quad (7)$$

$$g_{j+1}[k] = g_j[k] \uparrow 2 = \begin{cases} g_j[\frac{k}{2}], & k \text{ even} \\ 0, & k \text{ odd} \end{cases} \quad (8)$$

where $h_0[k] = h[k]$ and $g_0[k] = g[k]$. The wavelet coefficients of signal x are calculated with a cascade of discrete convolutions.

$$a_{j+1}[n] = \bar{h}_j * a_j[n], \quad (9)$$

$$d_{j+1}^1[n] = \bar{g}_j * a_j[n], \quad (10)$$

where $a_0 = x$, $j = 0, \dots, J - 1$.

In this paper, the wavelet coefficients (either approximation or detail) of signal x at level L are called $\overset{\psi_L}{W}x$, which is related to x as

$$\overset{\psi_L}{W}x[n] = \bar{\psi}_L * x[n], \quad (11)$$

where

$$\psi_L[n] = f_0 * \dots * f_{L-1}[n], \quad (12)$$

and f is either h or g .

Wavelet functions have a property that interferes with extracting moment invariants in the corresponding domain: their moments are zero up to a certain order which depends on the function, and are called vanishing moments.

Definition 6. *The wavelet function $\psi \in L^2(\mathbb{Z})$ has M_ψ vanishing moments if*

$$\int_{-\infty}^{+\infty} t^p \psi(t) dt = 0, \quad \text{for } p \leq M_\psi. \quad (13)$$

The number of vanishing moments of ψ is equal to the number of zeros of $\hat{\psi}(w)$ at $w = 0$ [12]. M_ψ depends on the number of zeros of $\hat{g}(w)$ at $w = 0$, M_g , and its repetition in obtaining ψ , N . Since the scaling filter is designed such that $\hat{h}(0) = \sqrt{2}$, it can be derived that $M_\psi = NM_g$. Therefore, if $\psi_L \in L^2(\mathbb{Z})$, and it consists of N wavelet filters, then it has $M_{\psi_L} = NM_g$ vanishing moments.

In the next section, the effect of vanishing moments and the way that it is dealt with is explained.

4 Blur Invariants

Having chosen the proper wavelet transform, the representation of blur and moments in the wavelet domain will be extracted.

4.1 Blur in the Wavelet Domain

The wavelet transform of a blurred signal with wavelet function ψ_L is

$$\overset{\psi_L}{W}y[n] = \bar{\psi}_L * y[n]. \quad (14)$$

Substituting y with its equivalent in (II) gives

$$\begin{aligned} \overset{\psi_L}{W}y[n] &= \bar{\psi}_L * b * x[n] \\ &= b * \bar{\psi}_L * x[n] = b * \overset{\psi_L}{W}x[n]. \end{aligned} \quad (15)$$

Eq.(15) implies that the wavelet transform of blurred signal y is the convolution of blur system b with the wavelet transform of original signal x .

4.2 Moments in the Wavelet Domain

The ordinary moment of order p of $\overset{\psi_L}{W}x$ is

$$\begin{aligned} m_p^{\overset{\psi_L}{W}x} &= \sum_n n^p \overset{\psi_L}{W}x[n] \\ &= \sum_n \sum_l n^p x[l] \psi_L[l-n] = \sum_l \sum_k (l-k)^p x[l] \psi_L[k] \\ &= \sum_{i=0}^p \sum_l \sum_k \binom{p}{i} (-1)^i l^{p-i} x[l] k^i \psi_L[k] = \sum_{i=0}^p \binom{p}{i} (-1)^i m_{p-i}^x m_i^{\psi_L}, \end{aligned} \quad (16)$$

where $l-n$ is substituted with k . If ψ_L has M_{ψ_L} vanishing moments, $m_i^{\psi_L}$ in

(16) will be zero for $i < M_{\psi_L}$. This forces the moments of $\overset{\psi_L}{W}x$ to also be zero for $p < M_{\psi_L}$. Considering this, (6) should be employed in order to calculate the central moments of the wavelet transform of signals. Therefore, the central

moment of order p of $\overset{\psi_L}{W}y$ is

$$\begin{aligned} \mu_p^{\overset{\psi_L}{W}y} &= \sum_n \left(n - \zeta^{\overset{\psi_L}{W}y} \right)^p \overset{\psi_L}{W}y[n] = \sum_n \sum_l \left(n - \zeta^{\overset{\psi_L}{W}y} \right)^p \overset{\psi_L}{W}x[l] b[n-l] \\ &= \sum_l \sum_k \left(\left(l - \zeta^{\overset{\psi_L}{W}x} \right) + (k - c^b) \right)^p \overset{\psi_L}{W}x[l] b[k] \end{aligned}$$

$$\begin{aligned}
&= \sum_{i=0}^p \sum_l \sum_k \binom{p}{i} \left(l - \zeta^{\psi_L x} \right)^i \zeta^{\psi_L x} [l] (k - c^b)^{p-i} b[k] \\
&= \sum_{i=0}^p \binom{p}{i} \mu_i^{\psi_L x} \mu_{p-i}^b,
\end{aligned} \tag{17}$$

where $l - n$ is substituted with k . Also, it is trivial to show that $\zeta^{\psi_L y} = \zeta^{\psi_L x} + c^b$. Since $\mu_i^{\psi_L x}$ is zero for $i < M_{\psi_L}$, p should be equal to or greater than M_{ψ_L} . Therefore, by defining $q + M_{\psi_L} = p$, $\dot{\mu}_q = \mu_p$, and $\binom{a}{b}_M = \binom{a+M}{b+M}$, (17) is modified to

$$\dot{\mu}_q^{\psi_L y} = \sum_{i=0}^q \binom{q}{i}_{M_{\psi_L}} \dot{\mu}_i^{\psi_L x} \mu_{q-i}^b. \tag{18}$$

4.3 Blur Invariants in the Wavelet Domain

It is clear from (18) that the central moments of a blurred signal are related to those of the original signal and the blur system. To have wavelet domain blur invariants based on these moments, it is required to find a combination of them such that the moments of the blur system are not present any longer.

Theorem 1. For $\zeta^{\psi_L x}$, which is the wavelet transform of x with wavelet function ψ_L , $C_q^{\zeta^{\psi_L x}}$ is invariant to symmetric and energy-preserving blur systems [10].

$$C_q^{\zeta^{\psi_L x}} = \begin{cases} \dot{\mu}_q^{\zeta^{\psi_L x}} - \frac{1}{\dot{\mu}_0^{\zeta^{\psi_L x}}} \sum_{l=0}^{\frac{q-1}{2}} \frac{\binom{q}{q-2l}_{M_{\psi_L}}}{\binom{2l}{0}_{M_{\psi_L}}} C_{q-2l}^{\zeta^{\psi_L x}} \dot{\mu}_{2l}^{\zeta^{\psi_L x}}, & l \text{ is odd} \\ 0, & l \text{ is even.} \end{cases} \tag{19}$$

Proof. Refer to [10] for the proof. \square

Theorem 1 is applicable when the blur system is energy-preserving. However, such systems are not always realistic. The next theorem eliminates this restriction.

Theorem 2. For $\zeta^{\psi_L x}$, which is the wavelet transform of x with wavelet function ψ_L , $D_q^{\zeta^{\psi_L x}}$ is invariant to symmetric blur systems.

$$D_q^{\zeta^{\psi_L x}} = \begin{cases} \dot{\nu}_q^{\zeta^{\psi_L x}} - \sum_{l=0}^{\frac{q-1}{2}} \frac{\binom{q}{q-2l}_{M_{\psi_L}}}{\binom{2l}{0}_{M_{\psi_L}}} D_{q-2l}^{\zeta^{\psi_L x}} \dot{\nu}_{2l}^{\zeta^{\psi_L x}}, & l \text{ is odd} \\ 0, & l \text{ is even,} \end{cases} \tag{20}$$

where $\dot{\nu}_q^x = \dot{\mu}_q^x / \dot{\mu}_0^x$.

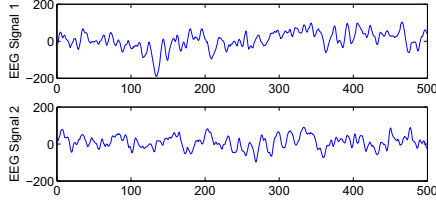


Fig. 1. The two EEG signals that are used in experiment 1

Proof. Assume \tilde{b} is a non-energy-preserving system. Such systems could be represented as

$$\tilde{b} = cb, \quad (21)$$

where c and b are a constant and an energy-preserving system, respectively. It is trivial to show that $\mu_q^{\tilde{b}} = c\mu_q^b$, and from here c can be found as the zero order moment of the non-energy-preserving system, $\mu_0^{\tilde{b}}$. Remember that the zero order moment of an energy-preserving system is equal to 1.

Assuming that y and z are both blurred versions of x by blur systems \tilde{b} and b , respectively, it can trivially be shown that $\nu^{\psi_L} y = \nu^{\psi_L} z$.

Without loss of generality, ν can be replaced in (19) to achieve blur invariants with no restriction on the energy of the system, given in (20). It should also be mention that since $\nu_0^{\psi_L} x$ becomes 1 for all signals, the term before the summation in (19) does not appear in (20) anymore. \square

5 Experimental Results

In this section, the modified invariants are evaluated in two different experiments: 1- EEG signals, 2- barcodes. To run the experiments, three well known wavelet filters are exploited: Coiflet of order 1, Daubechies of order 2, and Symlet of order 3 [2], which have 2, 2, and 3 vanishing moments, respectively. Also every experiment is carried out at different levels, which is indicated by mentioning the sequence of filters that are utilized. For example, hg means that the wavelet transform of signals are obtained by applying the lowpass filter followed by the highpass filter.

5.1 Experiment 1

Fig. 1 shows the two EEG signals that are used for this experiment. They are chosen from the database generated by Andrzejak et al. [1]. The signals are blurred by averaging on N neighborhoods, where N is 11 and 21, and energies of 0.7 and 0.4, respectively. Their wavelet transforms are obtained with wavelet filter Daubechies of order 2 at level ghh .

Table 1. Invariants of the EEG signals degraded with different blur systems. The wavelet filter is Daubechies of order two at level ghh . O.M. stands for the order of magnitude of the blur invariants. N is the number of neighborhoods in averaging. $N = 0$ refers to the original signal. The invariants of different orders do not change much by the variation of blur intensity and system energy, and they are sufficient to distinguish between the two signals.

$q/\text{O.M.}$		5/7	7/13	9/19	11/25
N/m_0^b					
Sig. 1	0	56.49	130.22	617.55	4317.66
	11/0.7	56.48	130.18	617.25	4314.94
	21/0.4	56.53	130.44	618.98	4330.91
Sig. 2	0	-0.41	0.21	-0.16	0.19
	11/0.7	-0.41	0.20	-0.16	0.19
	21/0.4	-0.40	0.20	-0.16	0.18

Table 1 represents the invariants of order 5 to 11 of the original signals and their blurred ones. It is clear that blur intensity and system energy changes do not affect the invariants significantly. Also, the two different signals can be perfectly distinguished from each other using the invariants.

5.2 Experiment 2

As a more challenging task, barcodes are used to evaluate the performance of the wavelet based blur invariants. In this experiment, the barcodes of two books are captured with a digital camera at different focuses, distances, and lighting conditions (Fig. 2). The first two factors introduce blur to the acquired signals, while the third one causes a change in the energy. The degradation level is so high in some of the images that it makes it impossible to distinguish between different bars with bare eyes. The images are then made of an equal width, changed from

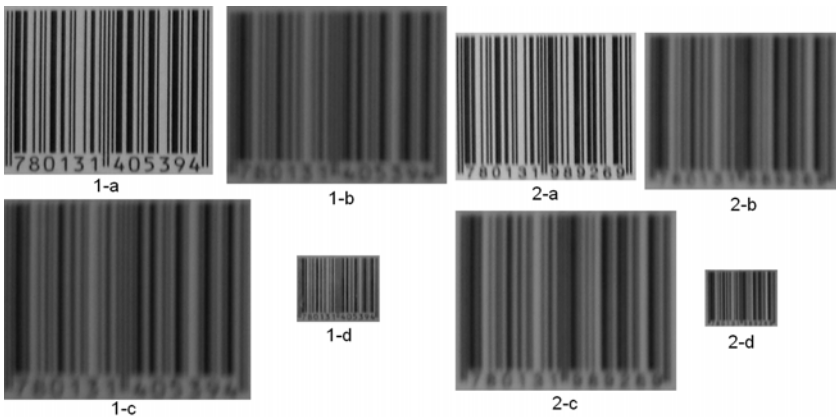


Fig. 2. Two barcodes at different focuses, scales, and lighting conditions

Table 2. Invariants of the central rows of the barcodes (the left half). The wavelet filter is Coiflet of order 1 at level gg . O.M. stands for the order of magnitude of the blur invariants. The invariants do not change drastically, although the images are degraded significantly by changes in focus, scaling, and lighting condition.

$q/O.M.$				
Barcode	5/5	7/10	9/15	11/20
1-a	-17.61	18.47	-30.80	73.97
1-b	-19.64	19.96	-32.08	74.23
1-c	-17.67	17.08	-26.60	59.82
1-d	-13.25	14.09	-23.24	55.02
2-a	-12.67	12.40	-19.51	44.36
2-b	-15.53	16.00	-25.63	59.30
2-c	-14.57	14.63	-23.07	52.67
2-d	-16.73	19.97	-35.41	89.53

Table 3. Invariants of the central rows of the barcodes (the right half). The wavelet filter is Symlet of order 3 at level gh . O.M. stands for the order of magnitude of the blur invariants. The variation within the invariants of similar barcodes is small, while there is a good difference between the invariants of the barcodes.

$q/O.M.$				
Barcode	5/5	7/10	9/15	11/20
1-a	2.16	-1.45	1.41	-1.98
1-b	1.75	-1.61	1.70	-2.47
1-c	2.13	-1.43	1.39	-1.94
1-d	2.75	-1.79	1.73	-2.42
2-a	-1.41	1.20	-1.35	2.13
2-b	-2.19	1.71	-1.87	2.89
2-c	-2.87	2.07	-2.22	3.44
2-d	-3.24	2.27	-2.43	3.74

RGB to greyscale, and their central rows are selected for comparison. As it can be seen in Fig. 2, the first 6 digits of these two barcodes are identical, however they are different on the second half. Therefore, the test is carried out on each half separately.

Tables 2 and 3 show the results of tests on the left and right sides, respectively. For the left side, the wavelet transforms are calculated with Coiflet of order 1 at level gg , and for the right side they are carried out with Symlet of order 3 at level gh . As it was expected, the results of the left half are very similar, and the two barcodes are easily distinguishable based on the results of the right half. Considering the different involved degrading factors, which are blur due to focus and scaling, lighting, and noise, the discrepancy among similar barcodes is not significant.

6 Conclusion

The wavelet blur invariants have been modified in this paper in order to make them available for applications where the blur system is not energy-preserving. Since most of the signals that are dealt with are discrete, the proper wavelet decomposition method is also chosen in order to keep the shift invariance property as well.

Two different experiments were chosen to evaluate the performance of the modified invariants. In the first experiment, EEG signals were blurred with blur systems of different intensities and energy levels, and in the second experiment, images of barcodes were acquired for a more challenging test. In both experiments, the invariants performed well showing very little variation due to changes in blur system effect and discriminating properly between different signals.

References

1. Andrzejak, R.G., Lehnertz, K., Mormann, F., Rieke, C., David, P., Elger, C.E.: Indications of nonlinear deterministic and finite-dimensional structures in time series of brain electrical activity: Dependence on recording region and brain state. *Phys. Rev. E* 64(6), 61907 (2001)
2. Daubechies, I.: *Ten Lectures on Wavelets*. Society for Industrial and Applied Math. (1992)
3. Flusser, J., Boldys, J., Zitova, B.: Moment forms invariant to rotation and blur in arbitrary number of dimensions. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25(2), 234–246 (2003)
4. Flusser, J., Suk, T.: Degraded image analysis: an invariant approach. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 20(6), 590–603 (1998)
5. Flusser, J., Suk, T.: Classification of degraded signals by the method of invariants. *Signal Processing* 60(2), 243–249 (1997)
6. Flusser, J., Suk, T., Saic, S.: Image features invariant with respect to blur. *Pattern Recognition* 28(11), 1723–1732 (1995)
7. Kingsbury, N.: Complex wavelets and shift invariance. *IEE Seminar on Time-scale and Time-Frequency Analysis and Applications (Ref. No. 2000/019)*, pp. 5/1–5/10 (2000)
8. Kundur, D., Hatzinakos, D.: Blind image deconvolution. *IEEE Signal Processing Magazine* 13(3), 43–64 (1996)
9. Lang, M., Guo, H., Odegard, J., Burrus, C., Wells, R.O.J.: Noise reduction using an undecimated discrete wavelet transform. *IEEE Signal Processing Letters* 3(1), 10–12 (1996)
10. Makaremi, I., Ahmadi, M.: Blur invariants: A novel representation in the wavelet domain. *Pattern Recognition* 43(12), 3950–3957 (2010)
11. Mallat, S.: Zero-crossings of a wavelet transform. *IEEE Transactions on Information Theory* 37(4), 1019–1033 (1991)
12. Mallat, S.: *A Wavelet Tour of Signal Processing, (Wavelet Analysis & Its Applications)*, 2nd edn. Academic Press, London (1999)

13. Ojansivu, V., Heikkilä, J.: Object recognition using frequency domain blur invariant features. In: Ersbøll, B.K., Pedersen, K.S. (eds.) SCIA 2007. LNCS, vol. 4522, pp. 243–252. Springer, Heidelberg (2007)
14. Shensa, M.J.: The discrete wavelet transform: wedding the a trous and mallat algorithms. *IEEE Transactions on Signal Processing* 40(10), 2464–2482 (1992)
15. Zhang, H., Shu, H., Han, G., Coatrieux, G., Luo, L., Coatrieux, J.: Blurred image recognition by legendre moment invariants. *IEEE Transactions on Image Processing* 19(3), 596–611 (2010)

A Super Resolution Algorithm to Improve the Hough Transform

Chunling Tu¹, Barend Jacobus van Wyk¹, Karim Djouani¹,
Yskandar Hamam¹, and Shengzhi Du²

¹ French South Africa Technical Institute(FSATI), Tshwane University of
Technology, Pretoria 0001, South Africa
tclchunling@gmail.com

² Department of Electrical and Mining Engineering, University of South Africa,
Pretoria 0002, South Africa

Abstract. This paper introduces a Super Resolution Hough Transform (SRHT) scheme to address the vote spreading, peak splitting and resolution limitation problems associated with the Hough Transform (HT). The theory underlying the generation of multiple HT data frames and the registration of cells obtained from multiple frames are discussed. Experiments show that the SRHT avoids peak splitting and successfully alleviates vote spreading and resolution limitations.

1 Introduction

The Hough Transform (HT) [1] is one of the most cited techniques for detecting straight lines and curves in gray level images. The basic idea of the HT is that a feature point (pixel) belonging to a straight line in the image space corresponds to a sinusoidal curve in the parameter space. The sinusoidal curves of all feature points on a straight line have a common intersection (i.e. the peak of accumulator matrix) which is used to detect the line in the parameter space. Different methods to improve the accuracy and resolution of HT, such as [2-13], were reported. Most of these focus on modifying the HT voting framework to increase accumulators to obtain higher accuracy and resolution. However, the discrete nature of the voting process causes peak generation problems in the HT space. It might split a peak into several peaks lying close to each other. It also spreads the peak to several cells around its ‘true’ position, causing the peak not to be distinct and hence limiting the accuracy of the HT, especially when disturbances and noise are present in the image space. An extreme case is that when the resolution is set so high that the peak will be too flat to be accurately detected.

In this paper a hybrid method, utilizing Super-Resolution (SR), is proposed to solve these problems. Multiple low-resolution (LR) HT data frames are generated to obtain new information and reconstruct a High Resolution (HR) HT data frame. To the authors’ knowledge, SR techniques have never before been used in this manner to improve the HT. SR image reconstructing refers to the process where a sequence of LR images is used to produce an HR image. Each LR

image must contain new information. The LR images in normal optical imaging are obtained via sub-pixel shifting of the camera. The main challenge related to this work is that the nature of the HT is such that traditional optical methods are not suitable for LR HT data frame generation and pixel (cell) registration. Furthermore, in many applications only one image is provided for object recognition. Multiple data frame generation and pixel (cell) registration related to the HT are discussed in this paper. Interpolation is used to reconstruct a HT data frame. The proposed Super Resolution Hough Transform (SRHT) scheme overcomes the vote spreading, peak splitting and resolution limitation problems associated with the Hough Transform (HT).

2 Hough Transform Problems

As shown in Fig. 1 the standard Hough Transform uses the following steps:

Step 1: Discretize the parameter space (HT space) into cells;

Step 2: Each feature point (x, y) contributes 1 vote to each cell on the sine curve

$$\rho = x \cos \theta + y \sin \theta, \quad (1)$$

i.e. feature points lying on the same straight line will vote to a common cell representing the straight in the HT space resulting in a peak in the accumulator matrix;

Step 3: The biggest peak (the cell getting the most votes) represents the most prominent straight line.

2.1 Peak Splitting

Shapiro [5] demonstrated the HT peak splitting problem via an example shown in Fig. 2, where the mapped image in the HT space of the longer straight line

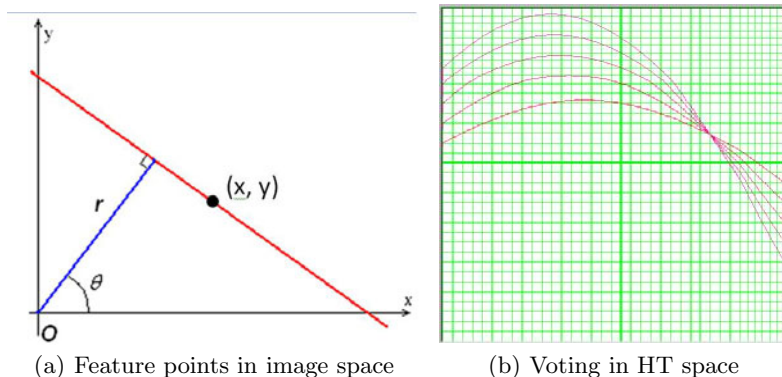


Fig. 1. Feature points vote to the cells lying on their sin curves

lies between two conjoint cells and hence are split into two peaks due to the rounding operation (shown in Fig. 2(b)), i.e. some feature points vote to one cell and the others vote to another cell. This causes the longer line to obtain a lower peak than the shorter one. It is obvious that this situation makes the detection process problematic. However, after the lines are moved by 0.5 pixel, the HT data becomes reasonable again as shown in Fig. 2(c), implying that the HT is not robust to small image shifts. Similar examples are not difficult

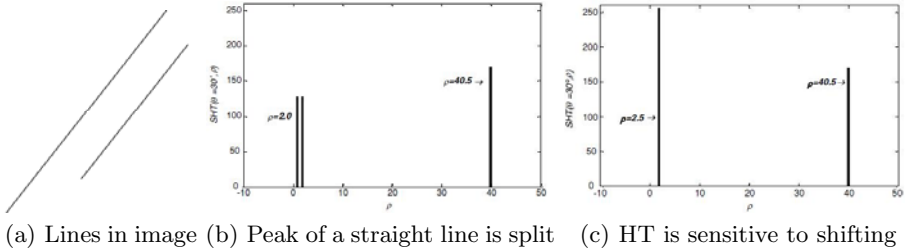


Fig. 2. Peak splitting in HT [5]

to construct. In fact, most cells share votes with conjoint cells because of the discretization and rounding operation during the voting process. In general the vote splitting/spreading problem cannot be avoided. It should be noted that not only the peaks might be split, the votes to each cell might also be unequally split.

2.2 Resolution Limitation

For the same image shown in Fig. 3, if the ρ -resolution is very high, the conjoint cells and the “true” cell get the same number of votes, implying that the HT cannot correctly locate straight lines in this situation. Vote spreading flattens the peaks when a very high resolution is required. Even peaks that are not flattened significantly will still affect detection accuracy and reliability. This paper proposes a solution to the open problem of how to obtain a reliable high resolution HT.

3 Using Super Resolution to Improve the HT

For the sake of simplicity, the idea of SR is demonstrated in Fig. 4 using a one dimensional case. Multiple LR images are generated by different imagers (or an imager using different viewpoints) to ensure new information exist in these LR images, i.e. no LR image can be obtained via other LR images. Then, according to the aliases of these LR images, their pixels are registered to a position in reference image. A HR image is then reconstructed by integrating the information of these LR images. By using the sensitivity of the HT to a shift in the images, multiple

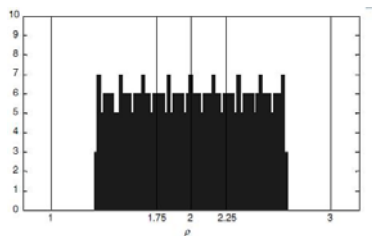


Fig. 3. Resolution Limitation in HT [5]

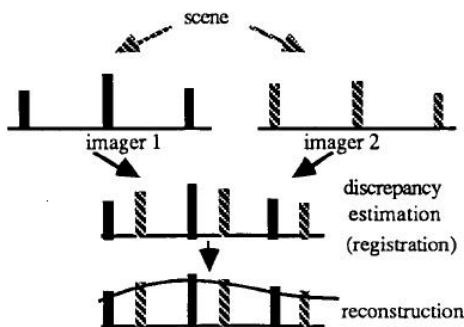


Fig. 4. The idea of SR [16]

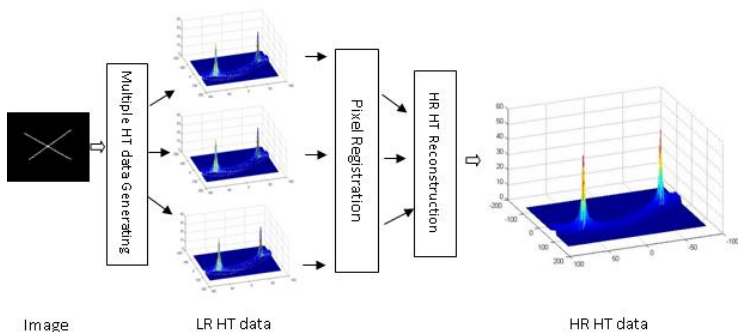


Fig. 5. The structure of SR-HT

LR HT data frames containing new information can be obtained via shifting. The resulting LR HT data frames are then used to construct a HR data frame using SR technologies.

The structure of the proposed SRHT method is demonstrated in Fig. 5. A given image is used to generate multiple LR HT data frames. These LR frames are then registered to a reference frame to reconstruct a HR frame.

3.1 Multiple HT Data Frame Generation

SR is based on the assumption that new information exists in the LR frames. Because of the information lost due to the rounding operation in optical imaging sensors, sub-pixel shifting ensures that each new LR frame contains new information. So, for normal optical images, cameras need to be aliased by fractional pixels as shown in Fig.6

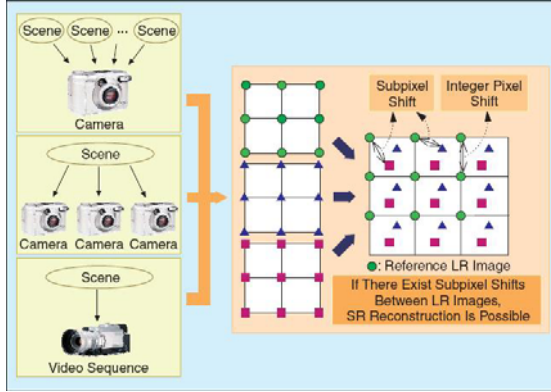


Fig. 6. Obtain multiple LR images via shifting camera by subpixel alias [15]

For this research, we assume that only one image is given, and present a method to generate multiple HT frames from this single image.

As previously mentioned, sub-pixel aliasing between frames are usually needed to generate images containing new information when dealing with digital optical images. The following will show what will happen in the HT space if the image is moved:

For feature point (x, y) we have

$$\rho = x \cos \theta + y \sin \theta. \tag{2}$$

After moving vertically by Δy we have

$$\rho' = x' \cos \theta' + y' \sin \theta' \tag{3}$$

where

$$\begin{aligned} x' &= x \\ y' &= y - \Delta y \\ \theta' &= \theta \end{aligned} \tag{4}$$

i.e.

$$\begin{aligned} \rho' &= x \cos \theta + (y - \Delta y) \sin \theta \\ &= x \cos \theta + y \sin \theta - \Delta y \sin \theta \\ &= \rho - \Delta y \sin \theta. \end{aligned} \tag{5}$$

If $\Delta y \sin \theta$ is not just equal to $n\Delta\rho$, i.e. new splitting ratios for point (x, y) appear around cell (θ, ρ) , then new information is generated during the shift. Similar results can be obtained if shifting horizontally or both vertically and horizontally. In fact, it is impossible for $\Delta y \sin \theta$ to simultaneously equal to $n\Delta\rho$ for all cells (θ, ρ) lying on the sine curve represented by eq. 2. Multiple HT data frames containing new information can therefore be obtained via shifting the image.

3.2 Pixel/Cell Registration

In digital optical images most pixels retain relatively strong neighbor relations between frames. The difference between frames are block based as shown in Fig. 7. Most macro blocks can find their corresponding blocks in other LR images. Motion estimation is a popular technique used in SR to register pixels in the reference frame. However, in HT data frames the difference between frames are

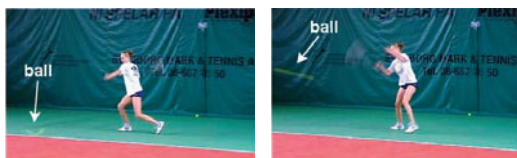


Fig. 7. Block based motion in optical images [14]

column-wise (shown in Fig. 8). Conjoint pixels (cells) in different columns will not be conjoint in other HT frames, i.e. relativity is broken, and the alias depends on both image shift and the position of the cell in the HT space. It is obvious that techniques like motion estimation is of no use in this situation.

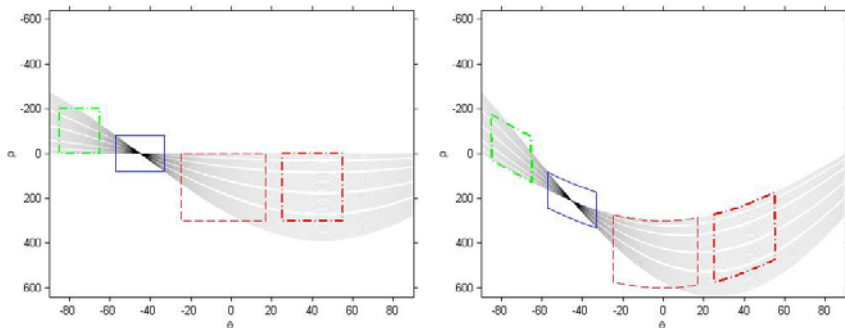


Fig. 8. Column based difference between HT data frames

We will now show how to register cells to the interpolating plane, i.e. calculate the alias of each cell in the reference frame, when multiple HT data frames

are available. As shown in eq. (5), the HT frame is aliased column-wisely after vertically shifting the given image. For the column corresponding to θ the alias is $-\Delta y \sin \theta$. So for a cell (θ', ρ') in the vertically shifted HT frame, its sub-sampling point in the reference frame is $(\theta', \rho' + \Delta y \sin \theta')$.

Similar to eq. (5), after horizontally moving the given image by Δx we have

$$\rho' = x' \cos \theta' + y' \sin \theta' \quad (6)$$

where

$$\begin{aligned} x' &= x - \Delta x \\ y' &= y \\ \theta' &= \theta \end{aligned} \quad (7)$$

i.e.

$$\begin{aligned} \rho' &= (x - \Delta x) \cos \theta + y \sin \theta \\ &= x \cos \theta + y \sin \theta - \Delta x \cos \theta \\ &= \rho - \Delta x \cos \theta. \end{aligned} \quad (8)$$

So for a cell (θ', ρ') in the horizontally shifted HT frame, its sub-sampling point in the reference frame is $(\theta', \rho' + \Delta x \cos \theta')$.

3.3 HR Reconstruction Using LR HT Frames

After registration, HR reconstruction is responsible for calculating the value of samples on the HR sampling points. The reconstruction methods for HT data are similar to those used for normal optical images. an interpolation method is used in this paper. Firstly a surface of the form $h = f(\theta, \rho)$ is fitted to the registered data in the non-uniformly spaced vectors (θ, ρ, h) where h represents the value of accumulators. Then this surface is interpolated at the points specified by HR $(\theta^{\text{HR}}, \rho^{\text{HR}})$ to produce h^{HR} .

4 Experiments

The image shown in Fig. 9(a) illustrates the proposed method:

4.1 The Improvement of Peak Splitting

After shifting the given image vertically and horizontally, multiple LR HT data frames are obtained as shown in Figs. 9(b), 9(c), and 9(d). In Fig. 9(d) the left peak is split into two. Furthermore, the LR HT data frames verified that the HT is very sensitive to a shift of the image. When the image is slightly shifted by only 1 or 2 pixels, their HT data frames have great differences in the value and width of peaks. This is because of the HT voting splitting/spreading problem, however, it provides new information in different frames and hence SR principles can be used to improve the HT performance. Fig. 9(e) shows the HT data frame obtained by SRHT where peak splitting is avoided and the width of peaks are also decreased, implying that the peaks are more distinct than the ones in LR HT data frames. HT-based image analysis methods will clearly benefit from these distinct peaks.

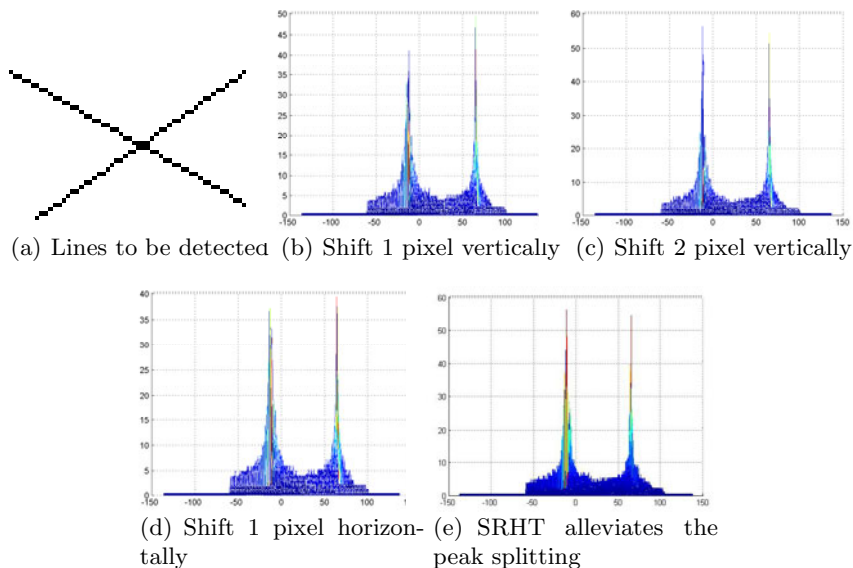


Fig. 9. Peak splitting and its solution by SRHT

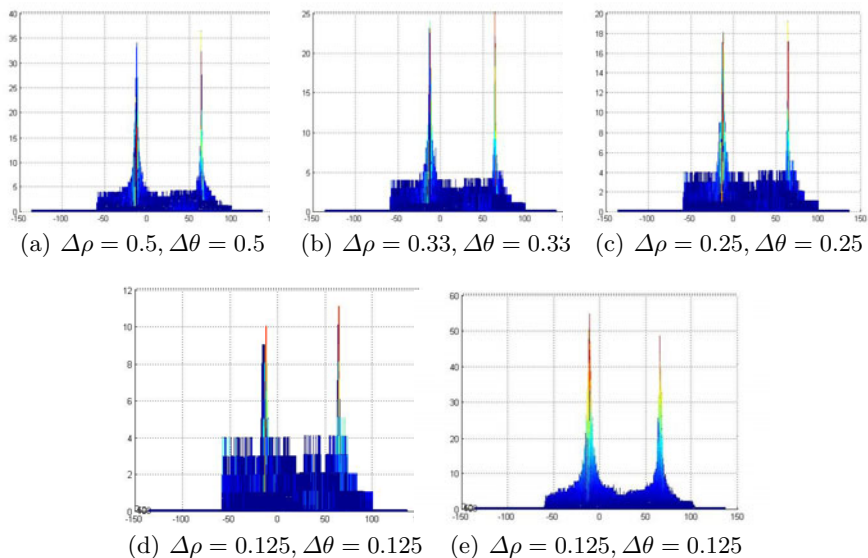


Fig. 10. The resolution limitation of HT and the improvement made by SR-HT

4.2 Improving the Resolution Limitation

When a very high resolution in the HT space is required, vote spreading becomes very prominent, i.e. the height of peaks becomes very low and the width of peaks become very wide as shown in Fig. 10(a)–10(d). This causes the HT to become unreliable and is the reason why high HT resolutions are normally avoided. Based on low resolution HT data frames obtained in Section 4.1, we obtain a high resolution HT data frame as shown in Fig. 10(e), where the peaks are very distinct even though the resolution is very high. As shown in Fig. 10(d) and Fig. 10(e), the HT data frames have the same resolutions, but the peaks of the latter are much more distinct than the former. The height of the peaks in Fig. 10(e) are almost as distinct as the ones obtained in the low resolution HT data frames shown in Section 4.1. This clearly demonstrates the ability of the proposed method to improve HT resolution limitations.

5 Conclusion

In this paper, SR technology was employed to improve the vote spreading, peak splitting and resolution limitation problems associated with the Hough Transform (HT). The theory underlying the generation of multiple HT data frames and the registration of cells obtained from multiple frames were discussed and a hybrid method, the Super Resolution Hough Transform (SRHT), was proposed. Experiments showed that the SRHT avoids peak splitting and successfully alleviates vote spreading and resolution limitations.

References

1. Hough, P.V.C.: A method and means for recognizing complex patterns. US Patent 3,069,654 (1962)
2. Duda, R.O., Hart, P.E.: Use of Hough transform to detect lines and curves in picture. *Communications of the ACM* 15(1), 11–15 (1972)
3. Song, J., Lyu, M.R.: A Hough transform based line recognition method utilizing both parameter space and image space. *Pattern Recognition* 38, 539–552 (2005)
4. Duan, H., Liu, X., Liu, H.: A nonuniform quantization of Hough space for the detection of straight line segments. In: *Proceedings of International Conference on Pervasive Computing and Applications (ICPCA 2007)*, pp. 216–220 (2007)
5. Shapiro, V.: Accuracy of the straight line Hough Transform: The non-voting approach. *Computer Vision and Image Understanding* 103, 1–21 (2006)
6. Walsh, D., Raftery, A.E.: Accurate and efficient curve detection in images: the importance sampling Hough transform. *Pattern Recognition* 35, 1421–1431 (2002)
7. Ching, Y.T.: Detecting line segments in an image - a new implementation for Hough Transform. *Pattern Recognition Letters* 22, 421–429 (2001)
8. Cha, J., Cofer, R.H., Kozaitis, S.P.: Extended Hough transform for linear feature detection. *Pattern Recognition* 39, 1034–1043 (2006)
9. Fernandes, L.A.F., Oliveira, M.M.: Real-time line detection through an improved Hough transform voting scheme. *Pattern Recognition* 41, 299–314 (2008)

10. Atiquzzaman, M., Akhtar, M.W.: Complete line segment description using the Hough transform. *Image Vision Comp.* 12(5), 267–273 (1994)
11. Atiquzzaman, M., Akhtar, M.W.: A robust Hough transform technique for complete line segment description. *Real-Time Imaging* 1(6), 419–426 (1995)
12. Du, S., van Wyk, B.J., Tu, C., Zhang, X.: An Improved Hough Transform Neighborhood Map for Straight Line Segments. *IEEE Trans. on Image Processing* 19(3) (2010)
13. Kamat, V., Ganesan, S.: A Robust Hough Transform Technique for Description of Multiple Line Segments in an Image. In: *Proceedings of 1998 International Conference on Image Processing (ICIP 1998)*, vol. 1, pp. 216–220 (1998)
14. Shechtman, E., Caspi, Y., Irani, M.: Space-Time Super-Resolution. *IEEE Transactions on Pattern Analysis And Machine Intelligence* 27(4), 531–545
15. Park, S.C., Park, M.K., Kang, M.G.: Super-Resolution Image Reconstruction: A Technical Overview. *IEEE Signal Processing Magazine* 21–36 (May 2003)
16. Komatsu, T., Aizawa, K., Igarashi, T., Saito, T.: Signal-processing based method for acquiring very high resolution images with multiple cameras and its theoretical analysis. In: *IEE Proceedings-I*, vol. 140(1), pp. 19–25 (February 1993)

Fusion of Multi-spectral Image Using Non-separable Additive Wavelets for High Spatial Resolution Enhancement

Bin Liu¹ and Weijie Liu²

¹ School of Mathematics and Computer Science, Hubei University, 430062, Wuhan, Hubei Province, China

² School of Computer, Wuhan University, 430079, Wuhan, Hubei Province, China
liub@hubu.edu.cn

Abstract. In order to solve the problems that the image fusion method based on separable discrete wavelet transform is lower in spatial resolution and there is block effect in fused image, a new multispectral image fusion method based on non-separable wavelets with compactly support, symmetry, orthogonality, and dilation matrix $[2,0;0,2]$ is proposed. A construction method of four channels 6×6 filter banks is presented. Using the low-pass filter constructed, multispectral images are fused. Three fusion methods called NAWS, NAWRGB and NAWL are proposed in the fusion of multispectral image and panchromatic image. Every fusion method presented outperforms the corresponding fusion method of the AWS, the AWRGB and the AWL in preserving high spatial resolution information respectively, and the higher spatial resolution fused image can be obtained. Of all fusion methods, the non-separable additive wavelet substitution (NAWS) method has the best performance in preserving higher spatial resolution information.

Keywords: Image fusion; Non-separable wavelets; Multispectral image; Panchromatic image.

1 Introduction

The fusion of multispectral image integrates the images which have higher spectral quality but lower spatial resolution and the panchromatic images with higher spatial resolution. It creates a new image which has better spectral information and higher spatial resolution. It is the hot technology of remote sensing image fusion, and has been widely used [1] [2] [3] [4] [5].

A number of approaches to pixel-level fusion have been proposed for merging multispectral image and panchromatic image [6] [7] [8] [9] [10]. The common procedures are intensity-hue-saturation mergers (IHS mergers or LHS mergers) [11], principal component analysis mergers (PCA mergers) [12], separable discrete wavelet transform (DWT mergers) [13]. All of these methods have their insufficiencies. The method of LHS transform can get high spatial resolution image, but the fused image may seriously lose the spectral information of the original MS image. The method of

PCA is adequate for the fusion of all wave bands multispectral images, and it can enhance the spatial resolution of the fused image, but it also reduces the spectral resolution and has a large amount of fusion computation. A fused image with good spectral information can be created by the separable discrete wavelet transform, but it is low in spatial resolution and there is block effect in the fused image.

In pixel-level image fusion using wavelets, reconstruction is a necessary step, so we deservedly have to construct the wavelets which have perfectly reconstruction performance. However, most wavelets applied in image processing in recently years are separable discrete wavelets generated from Daubechies one-dimensional wavelets via tensor product. Daubechies wavelets are not symmetric except Haar wavelet [14]. It is well known that a real value function has linear phase only when it is symmetric [15]. If a wavelet has not linear phase, then it will not have the properties of perfectly reconstruction. That is to say, in the separable discrete wavelets which have compact support and orthogonality, only the Haar wavelet has perfectly reconstruction performance, but Haar wavelet is too simple to keep good performance in lots of applications. So, how to construct the symmetrical wavelets that have good fusion performance is a key problem of wavelet application in image fusion. Biorthogonal wavelets are symmetrical, but it has not orthogonality, and when it is applied to image processing, information redundancy will be produced. Non-separable wavelets with symmetry will be constructed and applied to the field of image fusion in this paper.

Jorge Núñez proposed an additive wavelet “*á trous*” algorithm to fuse multispectral images and panchromatic images [16]. Its low-pass filter was generated from the one-dimensional B3 spline wavelet filters via tensor product. The proposed fusion method has better fusion effect in preserving spectral information. For extracting higher spatial resolution information from source panchromatic image, no quantitative analysis was made in the paper.

Non-separable wavelet is a new kind of wavelet developed in recently years [17] [18]. Compared to the separable wavelet, it has many good characteristics and can extract higher resolution information [19] [20]. Our group has studied the fusion methods of multispectral image and panchromatic image based on two channels non-separable wavelets [21] [22]. These methods have good fusion effect. Less computation amount is spent when images are decomposed and reconstructed. However, less information is obtained because only the diagonal line elements of the two channel filters are non-zero. The image information of the pixels whose positions are not in the diagonal line was lost.

This paper will spread the fusion method of multispectral image and panchromatic image which is based on separable additive wavelets to non-separable additive wavelets and explore its fusion performance.

2 Non-separable Orthogonal Wavelets with Compact Support and Filter Banks

When the sampling matrix is equal to $[2, 0; 0, 2]$, its determinant has an absolute value of 4. According to the theory of general two-dimensional wavelet transform, there are one scale function and three wavelet functions. Accordingly, there are four filters—a low-pass filter and three high-pass filters. Suppose H_0 and $H_i (i=1, 2, 3)$

are the low-pass filter and the high-pass filters of the decomposition respectively. H_0^* and H_i^* ($i=1,2,3$) are the low-pass filter and the high-pass filters of the reconstruction respectively. The key problem of constructing non-separable wavelets is to construct the non-separable wavelet low-pass filter and high-pass filters.

Suppose the dilation matrix of wavelet transform is $[2,0;0,2]$. According to the general constructing method of high dimensional wavelets with compactly support and orthogonality [23], we can construct the two-dimensional $2P \times 2P$ filter bank with compactly support, symmetry and orthogonality as formula (1).

$$(M_0(x, y), M_1(x, y), M_2(x, y), M_3(x, y)) = \frac{1}{4}(1, x, y, xy) \prod_{j=1}^K (U_j D(x^2, y^2) U_j^T) V \tag{1}$$

Where $x = e^{-i\omega_1}$, $y = e^{-i\omega_2}$, $M_i(x, y)$ ($i=0,1,2,3$) are the Fourier transform of H_i ($i=0,1,2,3$), and are also the frequency domain form of H_i ($i=0,1,2,3$). U_j ($j=1,2,\dots,K$) are a family of center-symmetric orthonormal matrices. $D(x, y) = \text{Diag}(1, x, y, xy)$. $V/2 = (V_0, V_1, V_2, V_3)/2$ is an orthonormal matrix. V_1, V_2, V_3 are 4×1 vectors. $V_0 = (1,1,1,1)^T$.

To seek 6×6 filter bank with symmetry, let $K=2$, constructing

$$A_j = \begin{pmatrix} 0 & 0 & \cos(\alpha_j) & -\sin(\alpha_j) \\ 0 & 0 & \sin(\alpha_j) & \cos(\alpha_j) \\ \cos(\beta_j) & -\sin(\beta_j) & 0 & 0 \\ \sin(\beta_j) & \cos(\beta_j) & 0 & 0 \end{pmatrix} (j=1,2) \qquad E = \begin{pmatrix} 1 & 0 & -1 & 0 \\ 0 & 1 & 0 & -1 \\ 0 & 1 & 0 & 1 \\ 1 & 0 & 1 & 0 \end{pmatrix}$$

And let

$$U_j = \frac{1}{2} \times E \times A_j \times E^T (j=1,2) \qquad V = \begin{pmatrix} 1 & -1 & 1 & -1 \\ 1 & 1 & 1 & 1 \\ 1 & 1 & -1 & -1 \\ 1 & -1 & -1 & 1 \end{pmatrix}$$

We can validate that: U_j ($j=1,2$) are center-symmetric orthonormal matrices, $V/2$ is an orthonormal matrix. Consequently, the filter bank H_0, H_1, H_2, H_3 which has compactly support, symmetry and orthonormality can be constructed. We have designed several groups of wavelet filter banks like this. From the experiments, we select $\alpha_1 = \pi/4, \alpha_2 = \pi/4, \beta_1 = \pi/3, \beta_2 = -\pi/3$, the low-pass filter of space domain form as formula (2) will be selected to fuse the images in the next section.

$$H_0 = \begin{pmatrix} 0.0534 & 0.0695 & 0.120 & -0.0924 & 0.000 & 0.000 \\ -0.00702 & 0.00915 & -0.0159 & -0.0122 & 0.000 & 0.000 \\ 0.0122 & 0.0159 & 0.188 & 0.188 & -0.120 & 0.0924 \\ 0.0924 & -0.120 & 0.188 & 0.188 & 0.0159 & 0.0122 \\ 0.000 & 0.000 & -0.0122 & -0.0159 & 0.00915 & -0.00702 \\ 0.000 & 0.000 & -0.0924 & 0.120 & 0.0695 & 0.0534 \end{pmatrix} \tag{2}$$

We can also validate that it is an orthonormal filter bank. Apparently H_0 is center-symmetric, H_1, H_2, H_3 are all center-symmetric filters, so this filter bank is a perfectly reconstructed filter bank with linear phase.

3 Fusion Algorithm

The algorithm steps are similar to “*à trous*” algorithm proposed in literature [16].

Firstly, decompose images using the two-dimensional non-separable low-pass filter as formula (2). Suppose p_0 is the source image, then

$$H_0(p_{i-1}) = p_i, w_i = p_{i-1} - p_i, (i=1, 2, \dots) \quad (3)$$

Where $w_i (i=1, 2, \dots)$ are wavelet planes, $p_i (i=0, 1, 2, \dots)$ are wavelet approximate components. Its reconstruction process is as formula (4).

$$p_0 = \sum_{i=1}^n w_i + p_r \quad (4)$$

Where p_r is the remnant image.

Secondly, three fusion modes were proposed similar to the literature [16] using the non-separable low-pass filter to replace the separable low-pass filter. In literature [16], three fusion modes were adopted to fuse multispectral and panchromatic images: (1) “Substitution method”: some of the wavelet planes of the multispectral image were substituted by the planes corresponding to the panchromatic image; (2) “Adding to the RGB Components”: adding the wavelet planes of the panchromatic image to R, G, and B directly; (3) “Adding to the Intensity Component”: adding the wavelet planes of the panchromatic decomposition to the intensity component of multispectral image. We call them as additive wavelet substitution (AWS), additive wavelet RGB (AWRGB) and additive wavelet L (AWL) respectively. Corresponding to the three fusion modes proposed in literature [16], the three fusion modes based on non-separable wavelet are used as follows: (1) non-separable additive wavelet substitution (NAWS); (2) non-separable additive wavelet RGB (NAWRGB); (3) non-separable additive wavelet L (NAWL).

4 Experimental Results Evaluation and Analysis

4.1 Experimental Results

In order to study the fusion performance of this kind of low-pass filters, we have designed a number of filter banks, and applied them to the fusion of multispectral and panchromatic images. We will show two experimental results at this time.

For the fusion of Quickbird satellite images, we select the panchromatic image whose spatial resolution is 0.61m. Figure 1(a) is the multi-band image whose spatial resolution is 2.44m. The three bands are the green band, the red band and the near



Fig. 1. Fusion of Quickbird MS image and PAN image. (a) Multispectral image; (b) Sub-image of MS image; (c) LHS fused sub-image; (d) DWT fused sub-image; (e) NAWs fused sub-image; (f) NAWRGB fused sub-image; (g) NAWL fused sub-image; (h) AWS fused sub-image; (i) AWRGB fused sub-image; (j) AWL fused sub-image; (k) Small image cut down from figure 1(e); (l) Small image cut down from figure 1(h).

infrared band. The bilinear interpolation method is used to re-sampling the multispectral image, and makes the pixel size of the sampling multispectral image be $0.61\text{m} \times 0.61\text{m}$ to accord with the resolution of the panchromatic image, and performs registration on panchromatic image and multispectral image. Because the image is too large, the sub-images have been cut out from the whole fused image to display the fusion effect clearly. Figure 1(b) is the sub-image of the source multispectral image.

Figure 1(e)-figure 1(g) are the fused sub-images of the proposed fusion method. In order to see the fusion effect clearly, we compare this method with the fusion methods based on LHS transform [11], DWT [13] and the separable additive wavelet *à trous* algorithm [16]. Figure 1(c) and figure 1(d) are the fused images of LHS method and DWT method respectively. Figure 1(h)-1(j) are the fused images of the AWS, AWRGB, AWL methods respectively. The wavelet function used in DWT is the db2 which is the second one of Daubechies series wavelets, and it has the same length as H_0 . The decomposition layers of the methods based on DWT, AWS, AWRGB, AWL and the proposed methods are 3, and the experiment is realized in the programming environment of MATLAB 7.5. In order to separate the low frequency and high frequency well, the filter of formula (2) was filtered by the 4×4 average value filter.

From the fusion effect of the proposed fusion methods, the spatial resolution and the spectral information of the MS images have been enhanced. That is, the results of the fusion contain structural details of the Pan image's higher spatial resolution and rich spectral information from the source multispectral images. The fused images of NAWS, NAWRGB and NAWL are clearer than the corresponding fused images of AWS, AWRGB and AWL, and all of them are clearer than the fused image by DWT method. From the visual effects, figure 1(e) has the highest spatial resolution. Figure 1(k) and figure 1(l) are the small images cut down from figure 1(e) and figure 1(h) respectively, figure 1(k) is clearer than figure 1(l) apparently. The fused image by LHS method has higher spatial resolution but its spectral information has degenerated seriously. The fused image based on DWT method preserves the spectral information of the multispectral image well, but it is vague and has lower spatial resolution.

Since NWAS has the best visual effect, we give another experiment to study the high-resolution performance of NWAS. For the fusion of LISS-3 images taken from IRS-P6 satellite, figure 2(a) and figure 2(b) are source images. Figure 2(a) is the LISS-3 panchromatic image which spatial resolution is 5.8m. Figure 2(b) is the LISS-3 multi-band images which spatial resolution is 23.5m. The three bands are the B2 band (green), the B3 band (red) and the B4 band (near infrared). This is a fire scene. Flame is burning on the upper right corner of the scene. The flame has been extinguished on lower-left corner of the scene. From color, the trace of burning and the flame could be seen.

Figure 2(f) is the fused image of NAWS. In order to see the fusion effects clearly, we also compare this method with the fusion methods based on IHS transform [11], DWT [13] and AWS [16]. Figure 2(c), figure 2(d) and figure 2(e) are the fused images of these three fusion methods respectively.

Comparing the visual effect of figure 2(f) with the fusion effects of figure 2(c), figure 2(d) and figure 2(e), the fused image of the proposed fusion method (NAWS) can preserve good spectral information and higher spatial resolution. The flame color

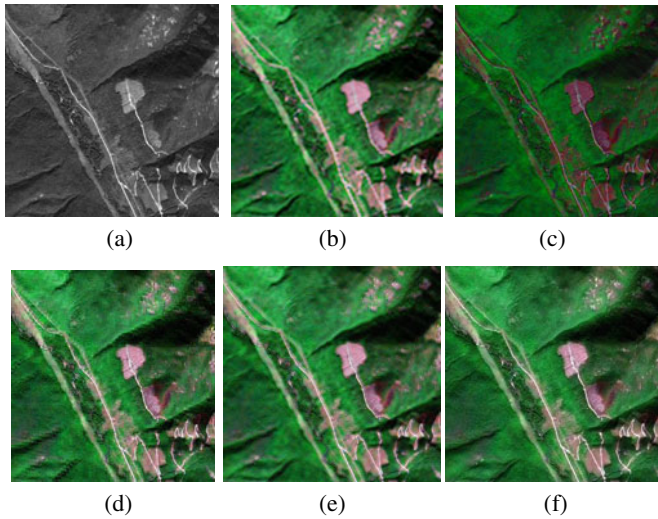


Fig. 2. Fusion of LISS-3 MS image and PAN image. (a) Original PAN image; (b) Original MS image; (c) IHS fused image; (d) DWT fused image; (e) AWS fused image; (f) NAWS fused image.

and the traces of burning of the scene are natural. It has no blocking artifact in the fused image. The fused images based on IHS fusion method (figure 2(c)) has higher spatial resolution, but the spectral information distorted badly. The color of fused image is deep. The fused image based on DWT has also good spectral information, but there are marked block effects in the ridge and other places. From the visual effects, figure 2(f) has the highest spatial resolution.

4.2 Performance Analysis of the Fused Result Images

The information entropy (IE) [10] [19] [20], the spatial frequency (SF) [24] and a kind of correlation coefficients (sCC) [3] [10] are used to measure the high spatial information the fused images preserve. The greater the entropy is, more information is contained. The entropy values of the Quickbird images fusion are listed in table 1. In literature [24], spatial frequency contains row frequency and column frequency as well as main diagonal frequency. The frequency in the spatial domain indicates the overall activity level in an image. The greater spatial frequency is, the clearer the images are, more spatial resolution information is contained. The spatial frequency of the different fusion method is presented in table 2. The high correlation coefficients between the fused filtered image and the Pan filtered image indicate that most of the spatial information of the Pan image was incorporated during the fusion process. These kinds of correlation coefficients of the fusion method proposed in this paper are presented in table 3.

Table 1. Entropy of the Quickbird fused images

<i>Fusion methods</i>		<i>R</i>	<i>G</i>	<i>B</i>
Initial MS Image		7.43	7.24	7.08
Additive Wavelet	AWS	7.50	7.34	7.20
	AWRGB	7.56	7.40	7.28
	AWL	7.49	7.31	7.17
Non-separable Additive Wavelet	NAWS	7.52	7.37	7.26
	NAWRGB	7.64	7.51	7.40
	NAWL	7.54	7.38	7.24
LHS		6.81	6.54	6.35
Separable wavelet		7.48	7.29	7.15

Table 2. Spatial frequency of the Quickbird fused images

<i>Fusion methods</i>		<i>R</i>	<i>G</i>	<i>B</i>
Initial MS Image		6.30	5.89	5.58
Additive Wavelet	AWS	14.75	14.64	14.54
	AWRGB	15.41	15.24	15.13
	AWL	10.44	10.17	10.01
Non-separable Additive Wavelet	NAWS	15.45	15.37	15.30
	NAWRGB	16.51	16.35	16.23
	NAWL	11.14	10.88	10.70
LHS		12.58	12.37	12.25
Separable wavelet		9.40	9.29	9.21

Table 3. Correlation coefficients between fused images and PAN images of Quickbird

<i>Fusion methods</i>		<i>R</i>	<i>G</i>	<i>B</i>
Additive Wavelet	AWS	0.988	0.988	0.988
	AWRGB	0.970	0.972	0.971
	AWL	0.915	0.919	0.917
Non-separable Additive Wavelet	NAWS	0.996	0.997	0.996
	NAWRGB	0.976	0.978	0.977
	NAWL	0.921	0.925	0.924
LHS		0.996	0.997	0.996
Separable wavelet		0.524	0.529	0.529

The results of table 1, table 2 and table 3 show that the fusion method based on non-separable additive wavelet performs better than the corresponding fusion method based on separable additive wavelet mentioned in literature [16] in preserving good spatial resolution information from the PAN image. NAWS has the highest spatial resolution in all the methods researched.

Table 4 lists the performance indices of NAWS fusion method for the LISS-3 fused images.

Table 4. Performance indices of the different fusion method for the LISS-3 fused images

	<i>LHS method</i>			<i>DWT method</i>			<i>AWS method</i>			<i>NAWS method</i>		
	R	G	B	R	G	B	R	G	B	R	G	B
IE	5.95	6.79	6.01	7.004	7.506	7.129	7.12	7.44	7.14	7.23	7.52	7.21
SF	10.31	10.80	10.71	15.35	14.80	15.14	17.22	17.23	17.35	18.19	17.83	18.03
sCC	0.934	0.972	0.986	0.937	0.976	0.968	0.977	0.986	0.987	0.994	0.993	0.996

Table 4 shows that NAWS fusion method has the highest spatial resolution.

5 Conclusion

This paper has presented a construction method of symmetric two-dimensional non-separable wavelet whose dilation matrix is $[2, 0; 0, 2]$, and has constructed non-separable wavelet low-pass filter, and has applied it in the fusion of multispectral image and panchromatic image. The proposed method has good fusion vision effect. From the objective performance indices, the fusion methods proposed outperforms the corresponding fusion method based on additive wavelet and the fusion method based on DWT in preserving the spatial resolution information, and the higher spatial resolution image can be obtained. Between the different non-separable additive wavelet based methods studied, the non-separable additive wavelet substitution (NAWS) method performs better in preserving higher spatial resolution information.

When using the methods proposed in this paper to fuse the SPOT-XS image and the SPOT-PAN image, the ETM+30m spatial resolution image and the ETM+PAN image, the SPOT high spatial resolution image and TM multispectral image, the same conclusion can be obtained.

Acknowledgments. This work was supported by the National Natural Science Foundation of China (61072126), and the key project of the Natural Science Foundation of Hubei Province (2009CDA133).

References

1. Pohl, C., Van Genderen, J.L.: Multisensor Image Fusion in Remote Sensing: Concepts, Methods and Applications. *International Journal of Remote Sensing* 19(5), 823–854 (1998)
2. Thomas, C., Ranchin, T., Wald, L., Chanussot, J.: Synthesis of Multispectral Images to High Spatial Resolution: A Critical Review of Fusion Methods Based on Remote Sensing Physics. *IEEE Transaction on Geoscience and Remote Sensing* 46(5), 1301–1312 (2008)
3. Zhou, J., Civco, D.L., Silander, J.A.: A Wavelet Transform Method to Merge Landsat TM and SPOT Panchromatic Data. *International Journal of Remote Sensing* 19(4), 743–757 (1998)
4. Heng, C., Weile, Z.: Fusion of IKONOS Satellite Imagery Using IHS Transform and Local Variation. *IEEE Geoscience and Remote Sensing Letters* 5(4), 653–657 (2008)
5. Moshoua, D., Bravao, C., Oberti, R., Westl, J., Bodria, L., McCartney, A., Ramon, H.: Plant Disease Detection Based on Data Fusion of Hyper-spectral and Multi-spectral Fluorescence Imaging Using Kohonen Maps. *Real-Time Imaging* 11(2), 75–83 (2005)

6. Yun, Z., Gang, H.: An IHS and Wavelet Integrated Approach to Improve Pan-sharpening Visual Quality of Natural Colour IKONOS and QuickBird Images. *Information Fusion* 6(3), 225–234 (2005)
7. Te-Ming, T., Shun-Chi, S., Hsuen-Chyun, S., Ping, S.H.: A New Look at HIS-like Image Fusion Methods. *Information Fusion* 2(3), 177–186 (2001)
8. Wang, Z., Ziou, D., Armenakis, C., Li, D., Li, Q.: A Comparative Analysis of Image Fusion Methods. *IEEE Transactions on Geoscience and Remote Sensing* 43(6), 1391–1402 (2005)
9. Ballester, C., Caselles, V., Igual, L., Verdera, J., Rougé, B.: A variational Model for p+xs Image Fusion. *International Journal of Computer Vision* 69(1), 43–58 (2006)
10. Myungjin, C., Rae, Y.K., Myeong, R.N., Hong, O.K.: Fusion of Multispectral and Panchromatic Satellite Images Using the Curvelet Transform. *IEEE Transaction on Geoscience and Remote Sensing Letters* 2(1), 136–140 (2005)
11. Chavez, P.S., Sides, S.C., Anderson, J.A.: Comparison of Three Different Methods to Merge Multiresolution and Multispectral Data: Landsat TM & SPOT Panchromatic. *Photogrammetric Engineering and Remote Sensing* 57(3), 295–303 (1991)
12. Sheffigara, V.K.: A Generalized Component Substitution Technique for Spatial Enhancement of Multispectral Image Using a High Resolution Data set. *Photogrammetric Engineering and Remote Sensing* 58(5), 561–567 (1992)
13. Yocky, D.A.: Image Merging and Data Fusion by Means of the Discrete Two-dimensional Wavelet Transform. *J. Opt. Soc. Amer.* 12(9), 1834–1841 (1995)
14. Daubechies, I.: *Ten Lecture on Wavelets*. Capital City Press, Philadelphia (1992)
15. Charles, K.C.: *An introduction to wavelets*. Academic Press, San Diego (1992)
16. Jorge, N., Xavier, O., Octavi, F., Albert, P., Vicenc, P., Roman, A.: Multiresolution-Based Image Fusion with Additive Wavelet Decomposition. *IEEE Transactions on Geoscience and Remote Sensing* 37(3), 1204–1211 (1999)
17. Kovačević, J., Vetterli, M.: Reconstruction Filter Bank and Wavelet Bases for \mathbb{R}^n . *IEEE Trans. on Information Theory* 38(2), 533–555 (1992)
18. Ayache, A.: Construction of Non-separable Dyadic Compactly Supported Orthonormal Wavelet Bases for $L^2(\mathbb{R}^2)$ of Arbitrarily High Regularity. *Revista Mathematica Iberoamericana* 15(1), 37–58 (1999)
19. Bin, L., Jiexiong, P.: Image Fusion Method Based on Non-separable Wavelet. *Machine Vision and Applications* 16(3), 189–196 (2005)
20. Bin, L., Jiexiong, P.: Image Fusion Method Based on Short Support Symmetric Non-separable Wavelet. *International Journal of Wavelets, Multiresolution, and Information Processing* 2(1), 87–98 (2004)
21. Bin, L., Jiexiong, P.: Multi-spectral Image Fusion Method Based on Two Channels Non-separable Wavelets. *Science in China Series F: Information Sciences* 51(12), 2022–2032 (2008)
22. Bin, L., Jiexiong, P.: Multi-spectral Image Fusion Based on Two Channels Non-separable Additive Wavelets. *Acta Optica Sinica* 27(8), 1419–1424 (2007) (in Chinese)
23. Qiuhui, C., Charles, A.M., Silong, P., Yuesheng, X.: Multivariate Filter Banks Having Matrix Factorizations. *SIAM J. Matrix Anal. Appl.* 25(2), 517–531 (2003)
24. Eskicioglu, A.M., Fisher, P.S.: Image Quality Measure and Their Performance. *IEEE Transaction on Communication* 43(12), 2959–2965 (1995)

A Class of Image Metrics Based on the Structural Similarity Quality Index

Dominique Brunet¹, Edward R. Vrscay¹, and Zhou Wang²

¹ Department of Applied Mathematics, Faculty of Mathematics, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

{dbrunet, ervrscay}@uwaterloo.ca

² Department of Electrical and Computer Engineering, Faculty of Engineering, University of Waterloo, Waterloo, Ontario, Canada N2L 3G1

zhouwang@ieee.org

Abstract. We derive mathematically a class of metrics for signals and images, considered as elements of \mathbf{R}^N , that are based upon the structural similarity (SSIM) index. The important feature of our construction is that we consider the two terms of the SSIM index, which are normally multiplied together to produce a scalar, as components of an ordered pair. Each of these terms is then used to produce a normalized metric, one of which operates on the means of the signals and the other of which operates on their zero-mean components. We then show that a suitable norm of an ordered pair of metrics defines a metric in \mathbf{R}^N .

Keywords: structural similarity index, normalized metrics, extended metrics, image quality assessment.

1 Introduction

Image quality assessment consists in modeling the perceptual fidelity between an original (ideal) image and a distorted version of it. The goal is not only to evaluate or compare the performance of image processing algorithms, but also to design an objective function to be optimized in order to develop better algorithms [1]. Traditionally, mean squared error (MSE) is used for this task, due to its simplicity and its many nice mathematical properties [1]. However, it is well known [1] that L^2 -based measures, e.g., mean squared error (MSE), are not necessarily good measures of visual quality.

Several image quality measures have been proposed in the literature as candidates to replace MSE [2]. While they generally outperform MSE in psycho-visual experiments, they are not known to share the mathematical properties of the MSE, making optimization very difficult to achieve. One concern is that these quality measures are not metrics in the strict mathematical sense since they do not satisfy the triangle inequality. As such, they are not amenable to standard procedures of mathematical analysis that may establish important properties, e.g., convergence, contractivity of operators.

An example of an application where these properties are important is collective sensing as described by Li in [3]. The main idea is to model an image as the

fixed point of a non-local operator, such as non-local means [4], BM3D [5] or a simplified version of non-local total variation [6]. One of the ideas of Li in [3] is to use an image representation that, contrary to cosine transforms and wavelets, is not based on a Hilbert space structure, but only on a metric space. Still, in his examples it was assumed implicitly that the metric used is the one associated with the L^2 -norm (i.e., MSE), which in fact is an inner product norm.

The structural similarity (SSIM) index [7] is an example of an image quality measure designed to provide better assessments of visual distortions between two images. The original formulation of the SSIM measure $S(\mathbf{x}, \mathbf{y})$ between two signals or images, $\mathbf{x}, \mathbf{y} \in \mathbf{R}_+^N$, involves a product of three terms, each of which measures a particular aspect of two images or image patches being compared, namely (i) the similarity of their mean values, (ii) the similarity between their contrasts and (iii) their correlation. The final two terms, however, can be collapsed into a single term. The resulting SSIM measure represents a combination of two pieces of information to produce a single number that characterizes the visual similarity of two image blocks. Such a procedure is known as *scalarization*. The question arises, however, whether it might be desirable to keep the two components, $S_1(\mathbf{x}, \mathbf{y})$ and $S_2(\mathbf{x}, \mathbf{y})$, of the SSIM separate, i.e., to treat the SSIM measure as a **vector**, an ordered pair, as opposed to a **scalar**. In this way, for example, their contributions could be weighted.

We show in this paper an example of a class of metrics for images derived from the SSIM index for which are associated neither norms nor inner products. This is done by first decomposing a signal \mathbf{x} into two orthogonal components, a one-dimensional space, \mathbf{R}_1^N , which involves only $\bar{\mathbf{x}}$, the mean of \mathbf{x} , and an $(N - 1)$ -dimensional space, \mathbf{R}_2^N , containing the zero-mean component of \mathbf{x} . We then show that if d_1 and d_2 are any two metrics on the spaces \mathbf{R}_1^N and \mathbf{R}_2^N , respectively, then the L^p norm of the ordered pair $\mathbf{d} = (d_1, d_2)$ is a metric on \mathbf{R}^N . Finally, we employ SSIM-based metrics for d_1 and d_2 in order to obtain our desired class of image metrics.

2 The Structural Similarity (SSIM) Quality Measure

In what follows, we let \mathbf{R}_+^N denote the space of non-negative N -dimensional signal/image blocks, i.e., $\mathbf{x} \in \mathbf{R}_+^N$ implies that $\mathbf{x} = (x_1, x_2, \dots, x_N)$, with $x_k \geq 0$, $1 \leq k \leq N$. We also consider the L^2 distance between two such signals $\mathbf{x}, \mathbf{y} \in \mathbf{R}_+^N$ to be the usual root mean squared error (RMSE), denoted as follows,

$$\|\mathbf{x} - \mathbf{y}\|_2 = \left[\frac{1}{N} \sum_{k=1}^N (x_k - y_k)^2 \right]^{1/2}. \quad (1)$$

The original definition of the SSIM measure between \mathbf{x} and \mathbf{y} is as follows,

$$S(\mathbf{x}, \mathbf{y}) = \left[\frac{2\bar{\mathbf{x}}\bar{\mathbf{y}} + \epsilon_1}{\bar{\mathbf{x}}^2 + \bar{\mathbf{y}}^2 + \epsilon_1} \right] \left[\frac{2s_{\mathbf{x}}s_{\mathbf{y}} + \epsilon_2}{s_{\mathbf{x}}^2 + s_{\mathbf{y}}^2 + \epsilon_2} \right] \left[\frac{s_{\mathbf{x}\mathbf{y}} + \epsilon_3}{s_{\mathbf{x}}s_{\mathbf{y}} + \epsilon_3} \right]. \quad (2)$$

where

$$\begin{aligned}\bar{\mathbf{x}} &= \frac{1}{N} \sum_{i=1}^N x_i, & \bar{\mathbf{y}} &= \frac{1}{N} \sum_{i=1}^N y_i, \\ s_{\mathbf{x}}^2 &= \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{\mathbf{x}})^2, & s_{\mathbf{y}}^2 &= \frac{1}{N-1} \sum_{i=1}^N (y_i - \bar{\mathbf{y}})^2, \\ s_{\mathbf{xy}} &= \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{\mathbf{x}})(y_i - \bar{\mathbf{y}}).\end{aligned}\quad (3)$$

The small positive constants $\epsilon_1, \epsilon_2 \ll 1$ are added for numerical stability along with an effort to accommodate the perception of the human visual system.

In the special case that $\epsilon_3 = \epsilon_2/2$, the above formula simplifies to the following product of two terms,

$$S(\mathbf{x}, \mathbf{y}) = S_1(\mathbf{x}, \mathbf{y})S_2(\mathbf{x}, \mathbf{y}) = \left[\frac{2\bar{\mathbf{x}}\bar{\mathbf{y}} + \epsilon_1}{\bar{\mathbf{x}}^2 + \bar{\mathbf{y}}^2 + \epsilon_1} \right] \left[\frac{2s_{\mathbf{xy}} + \epsilon_2}{s_{\mathbf{x}}^2 + s_{\mathbf{y}}^2 + \epsilon_2} \right], \quad (4)$$

It is this form of SSIM, which is frequently used in applications, that will be examined in this paper. The extension to the three-term formulation in (2), if desired, is straightforward.

The component S_1 in (4) measures the similarity of the mean values, $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ of, respectively, \mathbf{x} and \mathbf{y} . Its functional form was originally chosen in an effort to accommodate Weber's law of perception [7]. The component S_2 in (4) is a combination of the correlation and a measure of contrast distortion (similarity between the variances) between \mathbf{x} and \mathbf{y} . Its functional form follows the idea of divisive normalization [8].

Since we are working with signals in $\mathbf{x}, \mathbf{y} \in \mathbf{R}_+^N$, it follows that $0 \leq S_1 \leq 1$ and $S_1(\mathbf{x}, \mathbf{y}) = 1$ if and only if $\bar{\mathbf{x}} = \bar{\mathbf{y}}$. Note also that $-1 \leq S_2(\mathbf{x}, \mathbf{y}) \leq 1$ and $S_2 = 1$ if and only if $\mathbf{x} - \bar{\mathbf{x}} = \mathbf{y} - \bar{\mathbf{y}}$. It implies that $-1 \leq S(\mathbf{x}, \mathbf{y}) \leq 1$ and that, for non-negative signals, $S(\mathbf{x}, \mathbf{y}) = 1$ if and only if $\mathbf{x} = \mathbf{y}$. (A negative value of $S(\mathbf{x}, \mathbf{y})$ implies that \mathbf{x} and \mathbf{y} are negatively correlated.) This suggests that the function,

$$T(\mathbf{x}, \mathbf{y}) = 1 - S(\mathbf{x}, \mathbf{y}), \quad (5)$$

could act as some kind of distance function, since $\mathbf{x} = \mathbf{y}$ implies that $T(\mathbf{x}, \mathbf{y}) = 0$. Note also that $0 \leq T(\mathbf{x}, \mathbf{y}) \leq 2$.

We now examine the components, S_1 and S_2 in (4), in this way. For S_1 ,

$$\begin{aligned}1 - S_1(\mathbf{x}, \mathbf{y}) &= 1 - \frac{2\bar{\mathbf{x}}\bar{\mathbf{y}} + \epsilon_1}{\bar{\mathbf{x}}^2 + \bar{\mathbf{y}}^2 + \epsilon_1} \\ &= \frac{|\bar{\mathbf{x}} - \bar{\mathbf{y}}|^2}{\bar{\mathbf{x}}^2 + \bar{\mathbf{y}}^2 + \epsilon_1}.\end{aligned}\quad (6)$$

The RHS of (6) may be viewed as a *normalized* squared L^2 distance between the mean values $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$. For S_2 ,

$$1 - S_2(\mathbf{x}, \mathbf{y}) = 1 - \frac{2s_{\mathbf{xy}} + \epsilon_2}{s_{\mathbf{x}}^2 + s_{\mathbf{y}}^2 + \epsilon_2}$$

$$= \frac{s_{\mathbf{x}}^2 + s_{\mathbf{y}}^2 - 2s_{\mathbf{x}\mathbf{y}}}{s_{\mathbf{x}}^2 + s_{\mathbf{y}}^2 + \epsilon_2} . \quad (7)$$

In the special case $\bar{\mathbf{x}} = \bar{\mathbf{y}} = 0$,

$$1 - S_2(\mathbf{x}, \mathbf{y}) = \frac{\|x - y\|^2}{\|x\|^2 + \|y\|^2 + \frac{N-1}{N}\epsilon_2} , \quad (8)$$

which is also a *normalized* squared L^2 distance between \mathbf{x} and \mathbf{y} . Equations (6) and (8) suggest that it is natural to consider SSIM-based metrics which operate on a decomposition of signals into their means and zero-mean components. This will be done in the next section.

3 A Class of SSIM-Based Metrics

3.1 Orthogonal Decomposition of the Signal/Image Space

Here, we shall work in the space \mathbf{R}^N of N -dimensional signals/image blocks. We also let $\mathbf{R}_2^N \subset \mathbf{R}^N$ denote the $(N - 1)$ -dimensional subspace (hyperplane) of zero-mean signals, i.e.,

$$\mathbf{x} = (x_1, x_2, \dots, x_N) \in \mathbf{R}_2^N \Rightarrow \bar{\mathbf{x}} = 0 \text{ or } \sum_{k=1}^N x_k = 0 . \quad (9)$$

Finally, define the one-dimensional subspace $\mathbf{R}_1^N = \text{span}\{(1, 1, \dots, 1)\}$, i.e.,

$$\mathbf{R}_1^N = \{\mathbf{y} = (y_1, y_2, \dots, y_n) \mid \mathbf{y} = c(1, 1, \dots, 1) \text{ for some } c \in \mathbf{R}\} . \quad (10)$$

\mathbf{R}_1^N and \mathbf{R}_2^N are orthogonal complements of each other since $\mathbf{x} \in \mathbf{R}_2^N$ and $\mathbf{y} \in \mathbf{R}_1^N$ implies that

$$\langle \mathbf{x}, \mathbf{y} \rangle = \sum_{k=1}^N x_k y_k = 0 . \quad (11)$$

Moreover,

$$\mathbf{R}^N = \mathbf{R}_1^N \oplus \mathbf{R}_2^N . \quad (12)$$

We shall denote the orthogonal decomposition of an element $\mathbf{x} \in \mathbf{R}^N$ in terms of these two subspaces as follows,

$$\mathbf{x} = \mathbf{x}_1 + \mathbf{x}_2, \quad \mathbf{x}_1 \in \mathbf{R}_1^N, \quad \mathbf{x}_2 \in \mathbf{R}_2^N . \quad (13)$$

The component \mathbf{x}_1 is the projection of \mathbf{x} onto the subspace \mathbf{R}_1^N , i.e.,

$$\mathbf{x}_1 = \langle \mathbf{x}, \hat{\mathbf{e}}_1 \rangle \hat{\mathbf{e}}_1, \quad \text{where } \hat{\mathbf{e}}_1 = \frac{1}{\sqrt{N}}(1, 1, \dots, 1) . \quad (14)$$

Therefore,

$$\mathbf{x}_1 = (\bar{\mathbf{x}}, \bar{\mathbf{x}}, \dots, \bar{\mathbf{x}}) = \bar{\mathbf{x}}(1, 1, \dots, 1) . \quad (15)$$

where $\bar{\mathbf{x}}$ is the mean of \mathbf{x} defined in (3). It follows that the zero-mean component, \mathbf{x}_2 , of \mathbf{x} in \mathbf{R}_2^N is given by

$$\mathbf{x}_2 = \mathbf{x} - \mathbf{x}_1 . \quad (16)$$

3.2 A Class of Two-Dimensional Metrics

The next step is to consider metrics on these orthogonal spaces. Let d_1 be a metric on \mathbf{R} and d_2 a metric on \mathbf{R}^{N-1} . Then for any two elements $\mathbf{x}, \mathbf{y} \in \mathbf{R}^N$, define the corresponding ordered pair,

$$\mathbf{d} = (d_1(\bar{\mathbf{x}}, \bar{\mathbf{y}}), d_2(\mathbf{x}_2, \mathbf{y}_2)) \in \mathbf{R}^2. \quad (17)$$

It is clear that $\mathbf{x} = \mathbf{y}$ implies that $\mathbf{d} = \mathbf{0}$. The following result shows that \mathbf{d} can be used to define a metric on \mathbf{R}^N .

Theorem 1. *Let $\|\cdot\|$ be a norm in \mathbf{R}^2 that satisfies the following increasing property in \mathbf{R}_+^2 : For any $\mathbf{a} \in \mathbf{R}_+^2$ and any positive ordered pair $\mathbf{b} = (b_1, b_2)$, with $b_1, b_2 > 0$,*

$$\|\mathbf{a} + \mathbf{b}\| \geq \|\mathbf{a}\|. \quad (18)$$

Then for \mathbf{d} defined in (17),

$$d(\mathbf{x}, \mathbf{y}) := \|\mathbf{d}\| \quad (19)$$

is a metric in \mathbf{R}^N .

Note 1. This theorem can be generalized for a combination of M metrics on \mathbf{R}^N .

Before proving this theorem we state that for any $p \geq 1$, the L^p norm in \mathbf{R}^2 satisfies the above increasing property. It also applies to the case $p = \infty$, i.e., the L_∞ norm. This can be checked by using Taylor's Theorem for multivariable functions.

Proof. It is quite straightforward to show that $d(\mathbf{x}, \mathbf{y})$ in (19) satisfies following necessary properties of a metric:

1. $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{y}, \mathbf{x})$ (symmetry),
2. $d(\mathbf{x}, \mathbf{y}) \geq 0$ (positivity),
3. $d(\mathbf{x}, \mathbf{y}) = 0$ if and only if $\mathbf{x} = \mathbf{y}$ (strict positivity).

It remains to prove that $d(\mathbf{x}, \mathbf{y})$ satisfies the triangle inequality, i.e., for any $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbf{R}^N$,

$$d(\mathbf{x}, \mathbf{y}) \leq d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y}). \quad (20)$$

This result follows from the assumptions that d_1 and d_2 are metrics and that the $\|\cdot\|$ norm satisfies the increasing property:

$$\begin{aligned} d(\mathbf{x}, \mathbf{y}) &= \|(d_1(\bar{\mathbf{x}}, \bar{\mathbf{y}}), d_2(\mathbf{x}_2, \mathbf{y}_2))\| \\ &\leq \|(d_1(\bar{\mathbf{x}}, \bar{\mathbf{z}}) + d_1(\bar{\mathbf{z}}, \bar{\mathbf{y}}), d_2(\mathbf{x}_2, \mathbf{z}_2) + d_2(\mathbf{z}_2, \mathbf{y}_2))\| \\ &= \|(d_1(\bar{\mathbf{x}}, \bar{\mathbf{z}}), d_2(\mathbf{x}_2, \mathbf{z}_2)) + (d_1(\bar{\mathbf{z}}, \bar{\mathbf{y}}), d_2(\mathbf{z}_2, \mathbf{y}_2))\| \\ &\leq \|(d_1(\bar{\mathbf{x}}, \bar{\mathbf{z}}), d_2(\mathbf{x}_2, \mathbf{z}_2))\| + \|(d_1(\bar{\mathbf{z}}, \bar{\mathbf{y}}), d_2(\mathbf{z}_2, \mathbf{y}_2))\| \\ &= d(\mathbf{x}, \mathbf{z}) + d(\mathbf{z}, \mathbf{y}). \end{aligned} \quad (21)$$

□

Note 2. The increasing property in (18) also holds for suitably weighted L^p norms, e.g.,

$$\|\mathbf{x}\| = \left[\sum_{k=1}^N w_{k,p} |x_k|^p \right]^{1/p}, \quad (22)$$

where $w_{k,p} > 0$ for $1 \leq k \leq N$. But (18) does not hold for **all** norms. That being said, the validity for L^p and weighted norms is sufficient for most, if not all, practical purposes.

3.3 The Normalized Metric Relevant to SSIM

We now return to the results of (6) and (8) in order to construct a SSIM-based metric. The following result will be necessary.

Theorem 2. For $M \geq 1$, let $\|\cdot\|_2$ the L^2 norm in \mathbf{R}^M . Then for $\epsilon \geq 0$, $\bar{d}: \mathbf{R}^M \times \mathbf{R}^M \rightarrow \mathbf{R}$, given by

$$\bar{d}(\mathbf{x}, \mathbf{y}) = \begin{cases} \frac{\|\mathbf{x}-\mathbf{y}\|_2}{\sqrt{\|\mathbf{x}\|_2^2 + \|\mathbf{y}\|_2^2 + \epsilon}}, & (\mathbf{x}, \mathbf{y}) \neq (\mathbf{0}, \mathbf{0}), \\ 0, & \mathbf{x} = \mathbf{y} = \mathbf{0}, \end{cases} \quad (23)$$

is a metric.

This theorem was proved for the case $\epsilon = 0$ in [9]. The proof for the case $\epsilon > 0$ will appear elsewhere [10].

Note 3. The metric \bar{d} is an example of a *normalized metric*. The range of values assumed by \bar{d} is the bounded interval $[0, \sqrt{2}]$: $\bar{d}(\mathbf{x}, \mathbf{y}) = 0$ when $\mathbf{x} = \mathbf{y}$ and, for $\epsilon = 0$, $\bar{d}(\mathbf{x}, \mathbf{y}) = \sqrt{2}$ when $\mathbf{y} = -\mathbf{x}$.

Note 4. For every $\epsilon \geq 0$, $\bar{d}(\mathbf{x}, \mathbf{0})$ is not a norm, since $\bar{d}(\alpha\mathbf{x}, \mathbf{0}) \neq \alpha\bar{d}(\mathbf{x}, \mathbf{0})$ for any $\alpha > 0$.

Note 5. The following is an interesting property of this metric: In the case $\epsilon = 0$,

$$\bar{d}(\mathbf{x}, \mathbf{0}) = 1 \quad \text{for all } \mathbf{x} \in \mathbf{R}^M. \quad (24)$$

This implies that no sequences $\{\mathbf{x}_n\}$ can converge to $\mathbf{0}$ in this metric: Even if $\mathbf{x}_n \rightarrow \mathbf{0}$ in the metric defined by the \mathbf{R}^M norm $\|\cdot\|$, i.e. $\lim_{n \rightarrow \infty} \|\mathbf{x}_n - \mathbf{0}\| = 0$, it cannot converge to $\mathbf{0}$ in \bar{d} metric since $\lim_{n \rightarrow \infty} \bar{d}(\mathbf{x}_n, \mathbf{0}) = 1$. This is not a major problem since, in general, we are concerned only with non-zero signals.

Nevertheless, this nonconvergence of sequences to $\mathbf{0}$ in the \bar{d} metric disappears when $\epsilon > 0$. This parameter will, in fact, appear if we consider nonzero stability constants in the SSIM function of (4).

Note 6. Once again in the case $\epsilon = 0$, we have a scale invariance property: For any $\alpha \in \mathbf{R}$, $\bar{d}(\alpha\mathbf{x}, \alpha\mathbf{y}) = \bar{d}(\mathbf{x}, \mathbf{y})$, which is consistent with (24).

Unlike the L^2 case (Euclidean metric), the level sets associated with this metric are nonconcentric (hyper)spheres. To illustrate, we consider the simple \mathbf{R}^2 case with $\epsilon = 0$. Let $\mathbf{a} = (a_1, a_2)$ denote a reference point in \mathbf{R}^2 . The C -level set of the metric \bar{d} is the set of $\mathbf{x} = (x_1, x_2)$ for which $\bar{d}(\mathbf{x}, \mathbf{a}) = C$, where $C \in [0, \sqrt{2}]$, i.e.,

$$\frac{\|\mathbf{x} - \mathbf{a}\|_2}{\sqrt{\|\mathbf{x}\|_2^2 + \|\mathbf{a}\|_2^2}} = C \Rightarrow \|\mathbf{x} - \mathbf{a}\|_2^2 = C^2\|\mathbf{x}\|_2^2 + C^2\|\mathbf{a}\|_2^2. \quad (25)$$

After a little algebra, we found that the level sets may be classified into the following cases:

Case 1: $0 \leq C < 1$. For each C -value, the corresponding C -level set is composed of the points $\mathbf{x} = (x_1, x_2)$ that satisfy the equation,

$$\left[x_1 - \frac{a_1}{1 - C^2} \right]^2 + \left[x_2 - \frac{a_2}{1 - C^2} \right]^2 = \frac{C^2(2 - C^2)}{(1 - C^2)^2} (a_1^2 + a_2^2). \quad (26)$$

This is a circle centered at $\frac{1}{1 - C^2}(a_1, a_2)$ with radius $r = \frac{C\|\mathbf{a}\|_2}{1 - C^2} \sqrt{2 - C^2}$.

The centers of these circles lie on the line that extends from the origin $\mathbf{0}$ to the point \mathbf{a} . They start at \mathbf{a} ($C=0$) and travel outward to infinity as $C \rightarrow 1^-$.

Case 2: $C = 1$. The level set is the line $a_1x_1 + a_2x_2$ which contains the origin $(0, 0)$. This line is perpendicular to the line that supports the centers of the level sets in Case 1.

Case 3: $1 < C \leq \sqrt{2}$. For each C -value the corresponding C -level set is composed of the points $\mathbf{x} = (x_1, x_2)$ that satisfy the equation,

$$\left[x_1 + \frac{a_1}{C^2 - 1} \right]^2 + \left[x_2 + \frac{a_2}{C^2 - 1} \right]^2 = \frac{C^2(2 - C^2)}{(C^2 - 1)^2} (a_1^2 + a_2^2). \quad (27)$$

This is a circle centered at $\frac{1}{C^2 - 1}(-a_1, -a_2)$ with radius $r = \frac{C\|\mathbf{a}\|_2}{C^2 - 1} \sqrt{2 - C^2}$.

Their centers of these circles lie on the line that extends from the origin $\mathbf{0}$ to the point $-\mathbf{a}$. They are coming in from infinity ($C = \sqrt{2}$) and travel toward $-\mathbf{a}$ as $C \rightarrow \sqrt{2}$. At $C = \sqrt{2}$, the level set is the single point $-\mathbf{a}$.

In Fig. [1](#) are plotted some level sets associated with the point $\mathbf{a} = (1, 1)$.

3.4 Construction of the SSIM-Based Metric

We may now define the SSIM-based metric that results from the above constructions. The normalized metric \bar{d} will be used in each of the subspaces \mathbf{R}_1^N and \mathbf{R}_2^N defined in Sect. [3.1](#).

Given $\mathbf{x}, \mathbf{y} \in \mathbf{R}^N$, we now define the following vector of metrics,

$$\mathbf{d}(\mathbf{x}, \mathbf{y}) = (d_1(\bar{\mathbf{x}}, \bar{\mathbf{y}}), d_2(\mathbf{x}_2, \mathbf{y}_2)) \in \mathbf{R}^2, \quad (28)$$

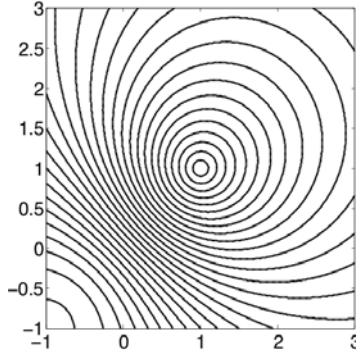


Fig. 1. Level sets $\bar{d}(\mathbf{x}, \mathbf{a}) = C$ about the reference point $\mathbf{a} = (1, 1)$ for $C = \frac{1}{20}k$, $1 \leq k \leq 28$, over the region $(x_1, x_2) \in [-1, 3] \times [-1, 3]$

where

$$d_1(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = \bar{d}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) = \frac{|\bar{\mathbf{x}} - \bar{\mathbf{y}}|}{\sqrt{\bar{\mathbf{x}}^2 + \bar{\mathbf{y}}^2 + \epsilon_1}}$$

$$d_2(\mathbf{x}_2, \mathbf{y}_2) = \bar{d}(\mathbf{x}_2, \mathbf{y}_2) = \frac{\|\mathbf{x}_2 - \mathbf{y}_2\|_2}{\sqrt{\|\mathbf{x}_2\|_2^2 + \|\mathbf{y}_2\|_2^2 + \frac{N-1}{N}\epsilon_2}} \quad , \quad (29)$$

The components, \mathbf{x}_2 and \mathbf{y}_2 of, respectively, \mathbf{x} and \mathbf{y} were defined in Sect. 3.1

In the particular case of two-dimensional signals, i.e., $N = 2$, which was illustrated in Fig. 1, we may view the d_1 metric as operating on the line $x_1 - x_2 = 0$ and the d_2 metric operator as operating on the orthogonal space $x_1 + x_2 = 0$ (zero-mean signals).

Now let $\|\cdot\|$ denote any norm in \mathbf{R}^2 satisfying the increasing property defined in Theorem 1. From that theorem, we have the resulting metric on \mathbf{R}^N :

$$D(\mathbf{x}, \mathbf{y}) = \|(\bar{d}(\bar{\mathbf{x}}, \bar{\mathbf{y}}), \bar{d}(\mathbf{x}_2, \mathbf{y}_2))\| \quad . \quad (30)$$

In the case that $\|\cdot\| = \|\cdot\|_p$, the weighted L^p norm on \mathbf{R}^2 , with $p \geq 1$, the metric is given explicitly as

$$D_p(\mathbf{x}, \mathbf{y}) = \|(\bar{d}(\bar{\mathbf{x}}, \bar{\mathbf{y}}), \bar{d}(\mathbf{x}_2, \mathbf{y}_2))\|_p$$

$$= (w_{1,p} [\bar{d}(\bar{\mathbf{x}}, \bar{\mathbf{y}})]^p + w_{2,p} [\bar{d}(\mathbf{x}_2, \mathbf{y}_2)]^p)^{1/p} \quad . \quad (31)$$

The cases $p = 1$ and $p = 2$ will probably be most relevant to standard image processing procedures:

$$D_1 = \bar{d}(\bar{\mathbf{x}}, \bar{\mathbf{y}}) + \bar{d}(\mathbf{x}_2, \mathbf{y}_2) \quad , \quad (32)$$

$$D_2 = \sqrt{\bar{d}^2(\bar{\mathbf{x}}, \bar{\mathbf{y}}) + \bar{d}^2(\mathbf{x}_2, \mathbf{y}_2)} \quad . \quad (33)$$

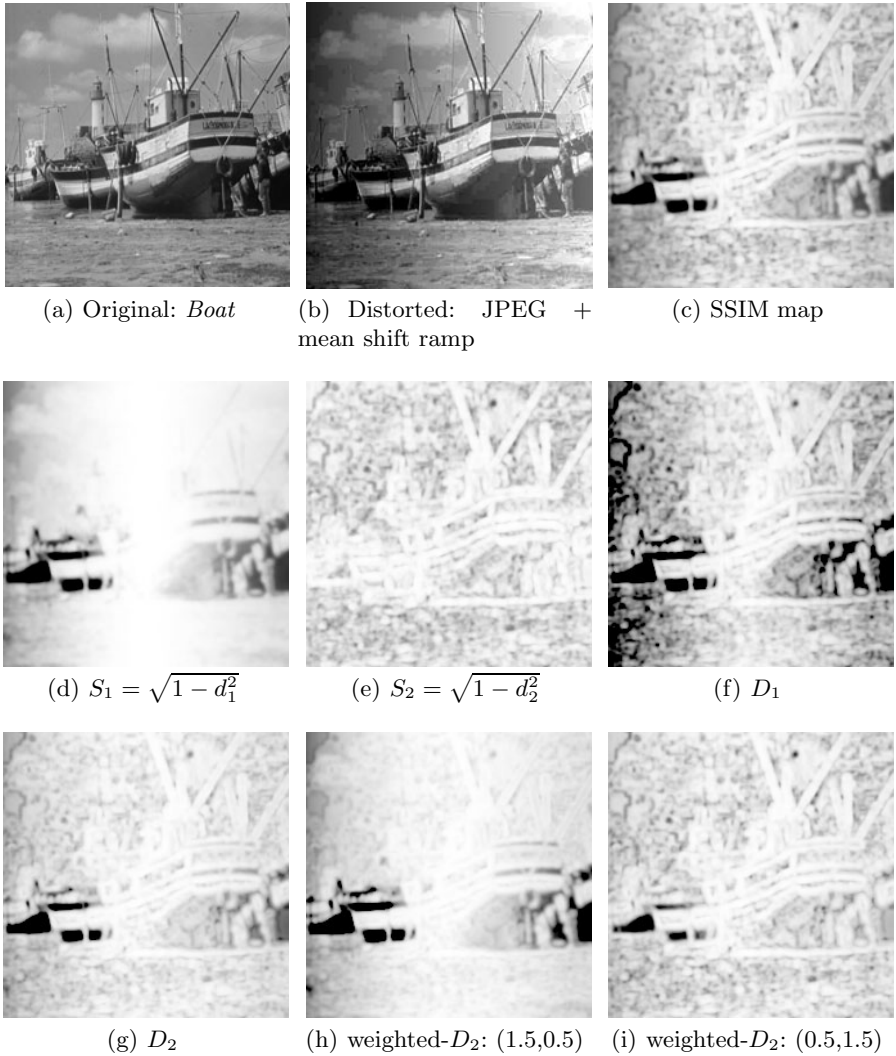


Fig. 2. (a) Original *Boat* image. (b) JPEG-compressed *Boat* image (quality factor 10/100) + horizontal mean shift ramp (from $-100/255$ to $+100/255$) (c) SSIM quality index map: Black=0, White=1. (d)-(e) S_1 and S_2 , the two components of the SSIM quality map (mean distortion and structural distortion) (f)-(i) SSIM-based metrics computed from different norms of $(d_1, d_2) = (\sqrt{1 - S_1}, \sqrt{1 - S_2})$: (f) D_1 , the L^1 -norm. (g) D_2 , the L^2 -norm. (h) Weighted L^2 -norm with $w_1 = 1.5$ and $w_2 = 0.5$. (i) Weighted L^2 -norm with $w_1 = 0.5$ and $w_2 = 1.5$. For all maps, a down-sampling was first performed and a sliding Gaussian window of $STD = 1.5$ pixels was used. The images (f)-(i) were rescaled with the formula $\sqrt{\max(0, 1 - D^2)}$ to look comparable to the SSIM map.

Note that in the special case $p = 2$ with unit weights, the above metric becomes the square root of the sum of the expressions in (6) and (7). Finally, the case $p = \infty$ may also be useful in some applications,

$$D_\infty(\mathbf{x}, \mathbf{y}) = \max\{\bar{d}(\bar{\mathbf{x}}, \bar{\mathbf{y}}), \bar{d}(\mathbf{x}_2, \mathbf{y}_2)\} . \quad (34)$$

By comparing the equation for D_2 with $\sqrt{1 - SSIM}$ we can understand their relationship:

$$\sqrt{1 - SSIM} = \sqrt{1 - (1 - d_1^2)(1 - d_2^2)} = \sqrt{d_1^2 + d_2^2 - d_1^2 d_2^2}. \quad (35)$$

D_2 may be viewed as a low order approximation of the SSIM index. In fact, most image distortions, e.g. JPEG and JPEG2000 compression, blur and zero-mean noise, preserve the mean. It implies that $d_1 = \bar{d}(\bar{\mathbf{x}}, \bar{\mathbf{y}})$ will be close to zero. Thus, D_2 is a very good approximation of SSIM for most of the distortions encountered in image processing. In fact, when the means are exactly matched, $D_2(\mathbf{x}, \mathbf{y}) = \sqrt{1 - SSIM}(\mathbf{x}, \mathbf{y})$.

Example 1. To offer some comparison between the new class of metrics and SSIM – and to show some of their limitations – we present an example involving a distortion of both the local structure and the local mean value. In Fig. 2 are shown several quality maps which compare the test image *Boat* (top left) with a JPEG compressed version (quality factor 10/100) to which was added a horizontal mean shift ramp from $-100/255$ to $+100/255$ (top middle). We see that all the different metrics detect the same error than the SSIM map, but none of them give exactly the same weight than SSIM for luminance distortion and structural distortion.

Psychovisual experiments will need to be performed to find the best parameters p and $w_{k,p}$ associated with these metrics. One of these metrics could be then used in image processing applications as optimization objective.

Acknowledgements. We gratefully acknowledge the generous support of this research by the Natural Sciences and Engineering Research Council of Canada (NSERC) in the forms of Discovery Grants (ERV, ZW), a Strategic Grant (ZW), a collaborative research and development (CRD) grant (ERV, ZW) and a Post-graduate Scholarship (DB). ZW would also like to acknowledge partial support by the Province of Ontario Ministry of Research and Innovation in the form of an Early Researcher Award.

References

1. Wang, Z., Bovik, A.C.: Mean squared error: Love it or leave it? A new look at signal fidelity measures. *IEEE Signal Processing Magazine* 26(1), 98–117 (2009)
2. Wang, Z., Bovik, A.C.: *Modern Image Quality Assessment*. Morgan & Claypool Publishers (2006)
3. Li, X.: Collective sensing: a fixed-point approach in the metric space. *SPIE Conf. on VCIP* (July 2010)

4. Buades, A., Coll, B., Morel, J.M.: A review of image denoising algorithms, with a new one. *Multiscale Modelling and Simulation* 4, 490–530 (2005)
5. Dabov, K., Foi, A., Katkovnik, V., Egiazarian, K.: Image denoising by sparse 3-D transform-domain collaborative filtering. *IEEE Trans. Image Processing* 16, 2080–2095 (2007)
6. Gilboa, G., Osher, S.: Nonlocal operators with applications to image processing. *Multiscale Modeling and Simulation* 7(3), 1005–1028 (2008)
7. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Processing* 13(4), 600–612 (2004)
8. Wainwright, M.J., Schwartz, O., Simoncelli, E.P.: Natural image statistics and divisive normalization: Modeling nonlinearity and adaptation in cortical neurons. In: Rao, R., et al. (eds.) *Probabilistic Models of the Brain: Perception and Neural Function*, pp. 203–222. MIT Press, Cambridge (2002)
9. Klamkin, M.S., Meir, A.: Ptolemy’s inequality, chordal metric, multiplicative metric. *Pacific J. Math.* 101(2), 389–392 (1982)
10. Brunet, D., Vrscay, E.R., Wang, Z.: On the mathematical properties of the structural similarity index (preprint). (March 2011), <http://www.math.uwaterloo.ca/~dbrunet/>

Structural Fidelity vs. Naturalness - Objective Assessment of Tone Mapped Images

Hojatollah Yeganeh and Zhou Wang

Department of Electrical and Computer Engineering, University of Waterloo,
Waterloo, Ontario, Canada N2L 3G1
yeganeh@uwaterloo.ca, zhouwang@ieee.org

Abstract. There has been an increasing number of tone mapping algorithms developed in recent years that can convert high dynamic range (HDR) to low dynamic range (LDR) images, so that they can be visualized on standard displays. Nevertheless, good quality evaluation criteria of tone mapped images are still lacking, without which, different tone mapping algorithms cannot be compared and there is no meaningful direction for improvement. Although subjective assessment methods provide useful references, they are expensive and time-consuming, and are difficult to be embedded into optimization frameworks. In this paper, we propose a novel objective assessment method that combines a multi-scale signal fidelity measure inspired by the structural similarity (SSIM) index and a naturalness measure based on statistics on the brightness of natural images. Validations using available subjective data show good correlations between the proposed measure and subjective rankings of LDR images created by existing tone mapping operators.

Keywords: image quality assessment, high dynamic range image, tone mapping, structural similarity, naturalness of images.

1 Introduction

The real world scenes exhibit a wide range of luminance variations. The dynamic range could be on the order of 10,000 to 1 from highlights to shadows [18]. High dynamic range (HDR) images allow us to capture greater luminance levels between its brightest and darkest regions than standard or low dynamic range (LDR) images. A common problem that is often encountered in practice is concerned about the visualization of HDR images – most display devices available to us have been designed to accommodate standard LDR images and cannot preserve all information contained in HDR images. In order to visualize HDR images using standard displays, a number of tone mapping algorithms have been proposed that convert HDR to LDR images, for example [15, 11, 8]. It should be noted that due to the dynamic range reduction, tone mapping operators (TMOs) unavoidably cause information loss. So the question is, having multiple TMOs a

hand, which TMO faithfully maintains the information in the HDR image, and which TMO produces the most natural-looking good quality LDR image?

Subjective evaluation is the most straightforward method to assess the performance of TMOs. In [7], perceptual evaluations were carried out for six TMOs with regard to similarity and preferences. Seven TMOs were compared in [22] using two architectural interior scene and fourteen subjects were asked to rate basic image attributes as well as naturalness of the LDR images. A more comprehensive subjective experiment was performed in [6], where ten observer were asked to rate LDR images generated by 14 TMOs in terms of brightness, contrast, details and colors, and also to rank the overall quality of the images. These subjective test data are useful references in studying tone mapping algorithms. However, subjective experiments tend to be time-consuming and expensive. In addition, the outcome from these experiments are difficult to be incorporated into the design and optimization of tone mapping algorithms. Moreover, subjective tests may not be able to provide a complete evaluation because subject cannot see all details of HDR images, whose information may be missing from the LDR images and the subjects may not be aware of the existence of the missing details.

The progress on objective assessment of tone mapped images has been quite limited. Typical objective image quality assessment approaches assume that the reference and test images have the same dynamic range [18], and thus are not applicable. A dynamic range independent approach was proposed in [3], where the authors used a visibility model of the human visual system (HVS) to compare pairs of HDR-LDR images and produce quality maps, which reflect the loss of visible features, the amplification of invisible features, and reversal of contrast polarity. These quality maps show good correlations with subjective classifications of image degradation types including blur, sharpening, contrast reversal, and no distortion. However, this method does not provide a single quality score for an entire image, making it impossible to be validated with subjective evaluations of overall image quality.

In this work, we aims to develop an objective quality assessment model for LDR images using their corresponding HDR images as references. Our model is composed of two components – structural fidelity measurement and naturalness assessment. The structural fidelity measure is inspired by the success of the structural similarity (SSIM) index [18], which has been shown to be well correlated with perceived image quality when tested using a number of large-scale subject-rated independent databases [19]. Its performance can be further improved when incorporated into a multi-scale framework [20]. However, SSIM or multi-scale SSIM models cannot be directly applied to compare images with different dynamic ranges. Our method is built upon multi-scale SSIM but is adapted to accommodate contrast comparisons across dynamic ranges. The naturalness assessment component in our approach is based upon brightness statistics of natural images. Although the model is simple, it appears to be useful and especially suited to the problem we are working with, where brightness mapping is an inevitable issue in the design of tone mapping algorithms.

2 Proposed Method

The invisibility of HDR reference image casts big challenges to objective quality assessment of tone mapped images. Because of the reduction of dynamic range, TMOs are deemed not to be able to preserve all information in HDR images, and human observers may not be aware of this. One of the most important factors in assessing TMOs is that how much structural information is preserved after tone mapping. In [21], we presented a novel approach to measure the structural fidelity between HDR and its tone mapped LDR images based on the philosophy of SSIM. However, this does not suffice to provide an overall quality evaluation of tone mapped images because an LDR image that maintains the structural information of the HDR image may not look natural, for example, in our study we observed some LDR images that well maintain the structural information in the HDR images look overly dark. Therefore, we would desire tone mapped images that achieve the best balance between two (sometimes competing) factors – structural fidelity preservation and high naturalness. Our quality assessment model is thus built upon these ingredients.

2.1 Structural Fidelity

Local Structural Fidelity Assessment. Our approach is derived from the philosophy behind the design of SSIM, which is based on the belief that the main purpose of human vision is to extract structural information from the visual scene, and thus perceived image distortion should be predictable by a measure of structural information loss. The original local SSIM definition includes a luminance, a contrast and a structure comparison components. Since the local luminance and contrast between HDR and LDR images are meant to be different, it does not make good sense to directly compare local luminance and contrast. Let x and y be two local image patches extracted from the HDR and LDR images respectively. Our local similarity measure is defined as

$$S_{\text{local}}(x, y) = \frac{2\sigma'_x\sigma'_y + C_1}{\sigma'^2_x + \sigma'^2_y + C_1} \cdot \frac{\sigma_{xy} + C_2}{\sigma_x\sigma_y + C_2}. \quad (1)$$

The second term is the structure comparison component as in SSIM, where σ_x , σ_y and σ_{xy} are the local standard deviations and cross correlation between the two patches in HDR and LDR images, respectively, and C_1 and C_2 are positive stabilizing constants. The modified local contrast comparison method is given in the first term, which is developed based on two considerations. First, the contrast difference between HDR and LDR image patches should not be penalized as long as their contrasts are both significant or both insignificant, as opposed to comparing images with the same dynamic range, where SSIM penalizes any change in contrast. Second, the algorithm should penalize the cases that the contrast is significant in one of the image patches, but insignificant in the other. The key issue here is to quantify the significance of local contrast. In order to do

this, we pass the local standard deviation through a nonlinear mapping function given by

$$\sigma' = \begin{cases} 0, & \sigma < T_1 \\ \frac{1}{2} \left\{ 1 + \cos \left[\frac{\pi}{T_2 - T_1} (\sigma - T_2) \right] \right\}, & T_1 \leq \sigma \leq T_2 \\ 1, & T_2 < \sigma, \end{cases} \quad (2)$$

where T_1 and T_2 are two threshold values that define the ranges of insignificant and significant contrasts, and a raised cosine function is employed to provide a smooth transition between the two ranges. Note that when two image patches are both significant (σ greater than T_2) or both insignificant (σ smaller than T_1), the first term of Eq. (II) equals 1, and thus the S_{local} measure is fully determined by the structure comparison component in Eq. (II).

Multi-scale Assessment. The local S_{local} measure described above is applied to an entire image using a sliding window approach across the image space, resulting in a quality map that indicates the quality variation across space.

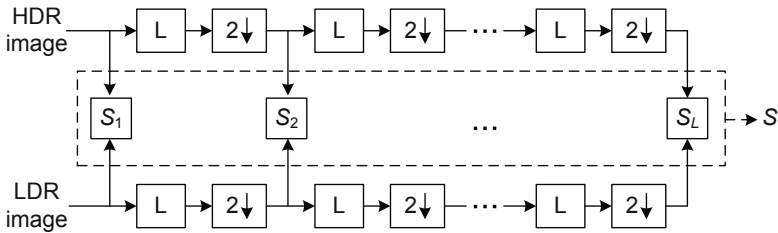


Fig. 1. Multi-scale framework of structural fidelity assessment method

The perceivability of image details also depends on the sampling density of the image signal, the distance from the image to the observer, the display resolution, and the perceptual capability of the observer’s visual system. In practice, the subjective evaluation of a given image varies with these parameters. A single-scale method as described in the previous section cannot capture such variations, and a multi-scale method is a convenient way to incorporate HVS features and image details at different resolutions. As in [20], we carry out signal fidelity assessment using a multi-scale structure depicted in Fig. II, where the images are iteratively low-pass filtered and downsampled, creating an image pyramid structure [4]. The local structural fidelity map is generated at each scale, and the map is then averaged to provide a single score for the scale by

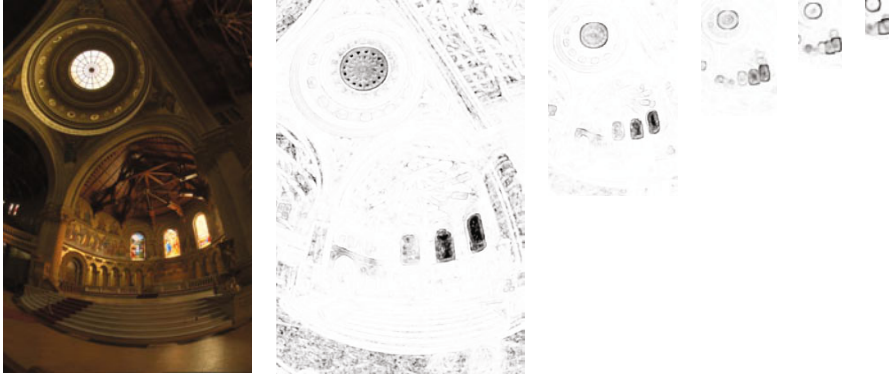
$$S_l = \frac{1}{N_l} \sum_{i=1}^{N_l} S_{\text{local}}(x_i, y_i), \quad (3)$$

where x_i and y_i are the i -th patches in the two images being compared, and N_l is the number of patches in the l -th scale. Fig. 2 shows examples of quality

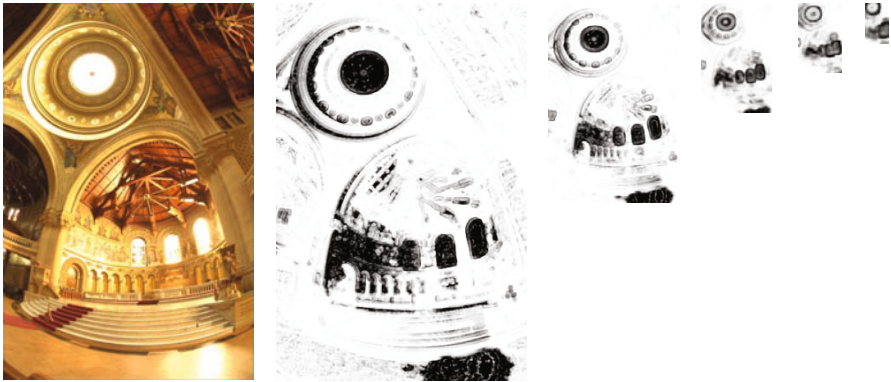
maps computed using the proposed multi-scale approach. Finally, the structural fidelity measures computed at each scale are combined to a multi-scale measure of the overall structural fidelity:

$$S = \prod_{l=1}^L S_l^{\beta_l}, \quad (4)$$

where L is the total number of scales and β_l is the weight assigned to the l -th scale.



(a) $S = 0.9288$ ($S_1 = 0.9371$; $S_2 = 0.9642$; $S_3 = 0.9524$; $S_4 = 0.9158$; $S_5 = 0.8286$)



(b) $S = 0.7980$ ($S_1 = 0.8419$; $S_2 = 0.8573$; $S_3 = 0.8330$; $S_4 = 0.7795$; $S_5 = 0.6361$)

Fig. 2. LDR images and their fidelity maps and scores in five scales. The images were created using Adobe Photoshop “Highlight compression” and “Exposure and Gamma” methods (not optimized for quality), respectively. The structural details of the brightest regions are missing in Image (b), but are more visible in Image (a). These are clearly reflected in the quality maps.

There are several parameters in the implementation of the multi-scale structural fidelity model. When computing S_{local} , we set $C_1 = 0.01$, $C_2 = 10$, $T_1 = 0.5$, and $T_2 = 4$, respectively. In our test, we find that the overall performance of our quality model is insensitive to these parameters within an order of magnitude, though fine tunings are yet to be performed through carefully designed psychophysical experiment. To create the fidelity map at each scale, we employ a Gaussian sliding window of size 11×11 with standard deviation 1.5. When combining the measures across scales, we set $L = 5$ and $\{\beta_l\} = \{0.0448, 0.2856, 0.3001, 0.2363, 0.1333\}$, which follows the psychophysical experiment results reported in [20]. To assess the quality of color images we first convert them from RGB color space to Yxy space and we apply the proposed structural fidelity measurement on luminance component Y only.

2.2 Naturalness

Tone mapping operators should be designed in a way that not only preserves structural information but also reproduces natural looking images. However, naturalness in general is a very subjective quantity and has not been clearly defined. A large literature has been dedicated to natural image statistics and their connections to biological vision. An excellent review can be found in [16]. Naturalness has also been studied in the context of subjective quality evaluation of tone mapped images. In [5], a subjective experiment was carried out and average correlation coefficients between image naturalness and different image attributes such as brightness, contrast, color reproduction, visibility and reproduction of details, are provided. The results show that among all attributes being tested, brightness and contrast have more correlation with perceived naturalness by subjects. This motivates us to build our naturalness model based on these two attributes. This choice may be oversimplifying in defining the general concept of image naturalness, but it captures the most important ingredients of naturalness that are related to the tone mapping evaluation problem we are trying to solve, where brightness mapping is an inevitable issue in all tone mapping operations.

Our method is built upon statistics of good-quality natural images. We gathered almost 3000 8bits/pixel natural images taken from many different scenes. These images are available at [12]. Figure 3 shows the histograms of the means and standard deviations of these images, which are useful measures that reflect the global luminance and contrast of images. We find that these histograms can be well fitted using a Gaussian and a Beta probability density functions, respectively, where the model parameters can be found by regression. The fitting curves are also shown in Fig. 3. Since brightness and contrast can be considered independent quantities in terms of both natural image statistics and biological computation [13], their joint probability density function would be the product of the two. Therefore, we define our naturalness measure as

$$N = \frac{1}{K} P_p P_c, \quad (5)$$

where K is a normalization factor given by $K = \max\{P_p P_c\}$, such that the naturalness measure is bounded between 0 and 1.

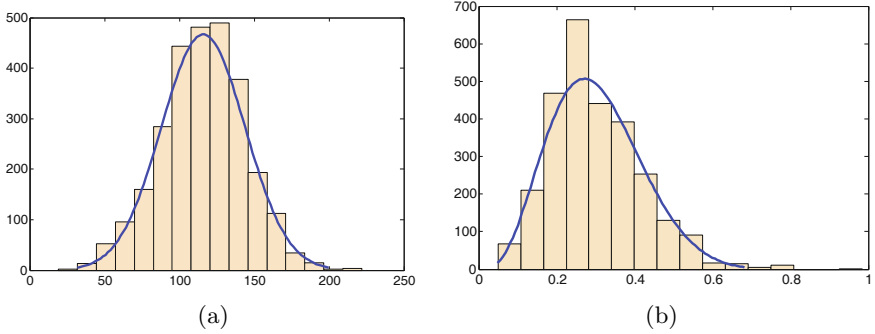


Fig. 3. Histograms of (a) means (fitted by Gaussian PDF) and (b) standard deviations (fitted by Beta PDF) of natural images

2.3 Quality Assessment Model

Given a tone mapped LDR image, we now have two available measurements, structural fidelity S and naturalness N , which are given by Eq. (4) and Eq. (5), respectively. These two quantities can be used individually or jointly as a 2D vector that characterizes different aspects of the quality of the LDR image. However, in most applications, users would prefer to have a single quality score of the image. Therefore, an overall quality evaluation that combines both quantities is desirable. In particular, we define the following 3-parameter function to combine the two components

$$Q = aS^\alpha + (1 - a)N^\beta, \quad (6)$$

where $0 \leq a \leq 1$ determines the relative weights assigned to the two components, and α and β defines the sensitivities of the two components, respectively. Since both S and N are upper-bounded by 1, this overall quality measure is also upper-bounded by 1. The parameters a , α and β , are left to be determined. In our implementation, they are tuned to best reflect subjective evaluations by utilizing machine learning techniques described next.

Machine Learning Process. The parameters in Eq. (6) can be learned from subjective quality evaluation data of tone mapped images. We were provided with subjectively ranked databases from the authors of [17], where the subjects were instructed to look at two LDR images at a time (produced by two different TMOs) and then choose the one with better quality. Two groups of studies have been carried out with such paired comparison approach. The first group of comparisons was conducted at Zhejiang University. 59 naive volunteers were invited to make the paired comparisons and fill the preference matrix. The second comparison was carried out by using Amazon Mechanical Turk, which is an online service for subjective evaluations. Each comparison task was assigned to 150 anonymous subjects. The database includes 6 folders, each of which contains images generated by 5 well-known TMOs, namely adaptive logarithmic mapping

[8], bilateral operator [9], uniform rational quantization [10], photoreceptor physiology [15] and exposure fusion [14]. The subjective ranking scores in each folder can then be computed using the preference matrix.

Finding the best parameters in Eq. (6) using subjective data is essentially a regression problem. The major difference from traditional regression problems is that here we are provided with relative ranking data between images only, but not quality scores associated with individual images. We developed an iterative method to learn the parameters. At each iteration, one pair of images is randomly selected from the database. If the model produce the correct order, then there is no change to the model parameters; Otherwise, each parameter is updated towards the direction of correcting the ranking error. To maintain the robustness of our approach, we carried out a cross validation process, where we divided the database into 6 folders and chose 5 as training set and the rest for testing. We repeat the same process 6 times, each with a different division between training and testing sets. Although each time ends up with a different set of parameters, they are fairly close to each other and result in the same ranking results. In the end, we fix $a = 0.8037$, $\alpha = 0.3958$ and $\beta = 0.8093$ as our final model parameters.

3 Validation

We used two independent subject-rated databases to test the proposed algorithm. The first is the database from [17] (which has also been used for training the parameters in Eq. (6)). We used leave-one-out cross-validation method described in the previous section to test our model. Table 1 shows the means and standard deviations of Kendall and Spearman rank order correlation coefficients between subjective rankings and our model predictions.

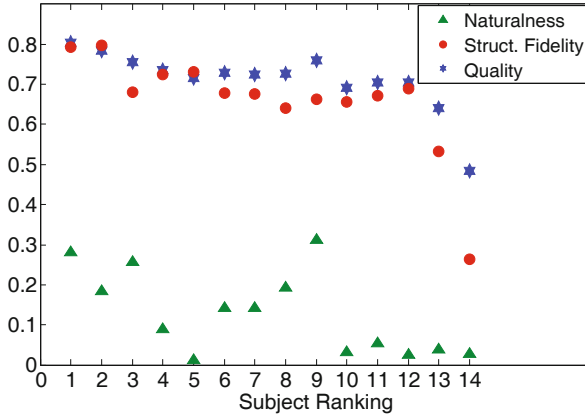
Table 1. Cross validation based on KRCC and SRCC using subjective data from [17]

	KRCC	SRCC
Mean	0.7333	0.8166
Std	0.2065	0.2136

The second database is from [6,12], where we utilized the overall quality rankings by 10 naive subjects of 14 tone mapped images. KRCC and SRCC between subjective rankings and our structural fidelity, naturalness and overall quality scores are given in Table 2. Fig. 4 shows the scatter plots of the results, where rank numbers 1 and 14 correspond to the best and worst quality images, respectively. It can be observed that the overall quality score generally agrees quite well with subjective rankings and is significantly better than using structural fidelity or naturalness measures alone. It is worth mentioning that the KRCC and SRCC values are even higher than those obtained in the training database, implying good generalization ability.

Table 2. KRCC and SRCC evaluations based on subjective data from [6,12]

	KRCC	SRCC
Structural Fidelity	0.6154	0.7967
Naturalness	0.4103	0.5606
Overall Quality	0.7692	0.8846

**Fig. 4.** Comparisons of subjective ranking versus structural fidelity, naturalness and overall quality scores using 14 tone mapped images from [6,12]

4 Conclusion

In this paper, we proposed an objective method to assess the quality of LDR images created from HDR images by tone mapping algorithms. The proposed approach is based on the combination of two measures, structural fidelity and naturalness. The structural fidelity measure follows the framework of the multi-scale SSIM approach to assess the structural information maintained after tone mapping operations. The naturalness criterion is designed by comparing with luminance statistics taken from natural scenes. Our experiments demonstrate that the proposed measure correlates well with subjective rankings of overall image quality. The proposed algorithm is computationally efficient and provides not only an overall quality score, but also multi-scale fidelity maps that indicate local structural variations across scale and space. As one of the initial attempts in objective assessment of tone-mapped images, the proposed method is quite promising and shows good potentials in the evaluation, design and optimization of tone mapping algorithms.

Acknowledgment

We would like to express our gratitude to the authors of [17] for providing us with their subjective test data at Zhejiang University and Amazon Mechanical Turk.

This research was supported in part by Natural Sciences and Engineering Research Council of Canada in the forms of Discovery, Strategic and CRD Grants, and by an Ontario Early Researcher Award, which are gratefully acknowledged.

References

1. <http://www-2.cs.cmu.edu/afs/cs/project/cil/www/v-images.html>
2. <http://www-staff.lboro.ac.uk/~cogs/datasets/UCID/ucid.html>
3. Aydm, T.O., Mantiuk, R., Myszkowski, K., Seidel, H.: Dynamic range independent image quality assessment. In: SIGGRAPH 2008: International Conference on Computer Graphics and Interactive Techniques, ACM SIGGRAPH (2008)
4. Burt, P.J., Adelson, E.H.: The Laplacian pyramid as a compact image code. *IEEE Trans. Communications* 31, 532–540 (1983)
5. Čadík, M., Slavík, P.: The naturalness of reproduced high dynamic range images. In: *IV 2005: Proceedings of the Ninth International Conference on Information Visualisation*, pp. 920–925. IEEE Computer Society, Washington, DC, USA (2005)
6. Čadík, M., Wimmer, M., Neumann, L., Artusi, A.: Image attributes and quality for evaluation of tone mapping operators. In: *Proceedings of the 14th Pacific Conference on Computer Graphics and Applications*, pp. 35–44. National Taiwan University Press, Taipei (2006)
7. Drago, F., Martens, W.L., Myszkowski, K., Seidel, H.P.: Perceptual evaluation of tone mapping operators. In: *Proc. of the SIGGRAPH Conf. Sketches and Applications* (2003)
8. Drago, F., Myszkowski, K., Annen, T., Chiba, N.: Adaptive logarithmic mapping for displaying high contrast scenes. *Computer Graphics Forum* 22(3), 419–426 (2003)
9. Durand, F., Dorsey, J.: Fast bilateral filtering for the display of high-dynamic-range images. *ACM Transactions on Graphics* 21, 257–266 (2002)
10. Fattal, R., Lischinski, D., Werman, M.: Gradient domain high dynamic range compression. In: *Proceedings of the 29th Annual Conference on Computer Graphics and Interactive Techniques, SIGGRAPH 2002*, pp. 249–256 (2002)
11. Ward Larson, G., Rushmeier, H., Piatko, C.: A visibility matching tone reproduction operator for high dynamic range scenes. *IEEE Transactions on Visualization and Computer Graphics* 3(4), 291–306 (1997)
12. Cadik, M., et al.: Evaluation of tone mapping operators, <http://www.cgg.cvut.cz/members/cadikm/tmo>
13. Mante, V., Frazor, R., Bonin, V., Geisler, W., Carandini, M.: Independence of luminance and contrast in natural scenes and in the early visual system. *Nature Neuroscience* 8(12), 1690–1697 (2005)
14. Mertens, T., Kautz, J., Van Reeth, F.: Exposure fusion. In: *Proceedings - Pacific Conference on Computer Graphics and Applications*, pp. 382–390 (2007), Cited By (since 1996): 8
15. Reinhard, E., Stark, M., Shirley, P., Ferwerda, J.: Photographic tone reproduction for digital images. In: *Proc. of 29th Annual Conference on Computer Graphics and Interactive Techniques. ACM SIGGRAPH*, vol. 21, pp. 267–276 (2002)
16. Simoncelli, E.P., Olshausen, B.A.: Natural image statistics and neural representation. *Annual Review of Neuroscience* 24 (2001)
17. Song, M., Tao, D., Chen, C., Bu, J., Luo, J., Zhang, C.: Exposure fusion using a probabilistic model. *IEEE Transactions on Image Processing* (submitted 2011)

18. Wang, Z., Bovik, A.C., Sheikh, H.R., Simoncelli, E.P.: Image quality assessment: From error visibility to structural similarity. *IEEE Trans. Image Proc.* 13, 35–44 (2004)
19. Wang, Z., Li, Q.: Information content weighting for perceptual image quality assessment. To appear in *IEEE Trans. Image Proc* (2011)
20. Wang, Z., Simoncelli, E.P., Bovik, A.C.: Multi-scale structural similarity for image quality assessment. In: *Proc. of 37th Asilomar Conf. Signals, Systems and Computers* (2003)
21. Yeganeh, H., Wang, Z.: Objective assessment of tone mapping algorithms. In: *Proc. IEEE Int. Conf. Image Proc.* (2010)
22. Yoshida, A., Blanz, V., Myszkowski, K., Seidel, H.: Perceptual evaluation of tone mapping operators with real-world scenes. In: *Human Vision and Electronic Imagin X*, SPIE, vol. 5666, pp. 192–203 (2005)

Learning Sparse Features On-Line for Image Classification

Ziming Zhang^{1,*}, Jiawei Huang², and Ze-Nian Li²

¹ School of Technology, Oxford Brookes University, Oxford, UK
ziming.zhang@brookes.ac.uk

² School of Computing Science, Simon Fraser University, B.C., Canada
{jha48,li}@cs.sfu.ca

Abstract. In this paper, we propose an efficient sparse feature on-line learning approach for image classification. A large-margin formulation solved by linear programming is adopted to learn sparse features on the max-similarity based image representation. The margins between the training images and the query images can be directly utilized for classification by the Naive-Bayes or the K Nearest Neighbor category classifier. Balancing between efficiency and classification accuracy is the most attractive characteristic of our approach. Efficiency lies in its on-line sparsity learning algorithm and direct usage of margins, while accuracy depends on the discriminative power of selected sparse features with their weights. We test our approach using much fewer features on Caltech-101 and Scene-15 datasets and our classification results are comparable to the state-of-the-art.

Keywords: Sparse feature selection, On-line learning, Image classification, Information retrieval.

1 Introduction

Efficiency in both training and classification phases is one of the most important characters for a successful image classification system. In this paper, we intend to design an on-line approach, where models are learned only using the data at hand regardless of the new data in future, so that it can achieve a good balance between efficiency and classification performance.

Imagine that you need to pick up a stranger at the airport and at hand you only have one of his photos. A natural way to find your stranger is to check everyone with the photo in eyes, nose, mouth, and/or other features. In this scenario, a simple query-template matching process is accomplished, which involves three different successive stages: 1. A template (*e.g.* the photo) needs to be defined first; 2. Then the matching process is performed between the template and the queries (*e.g.* all the people) based on some feature similarities (*e.g.* eyes, nose, mouth); 3. Finally, a decision is made about whether they are the same.

* This work was done when the author was at Simon Fraser University.

This simple template matching scheme has following attractive properties. First, it may be closer to the human perception process, as stated in [1] that the target category is defined by the similarities to the templates in the category rather than the lists of features. Second, during the matching process, only a small portion of characteristic features in the templates are needed. Third, it allows each template to make a decision independently no matter how many templates are at hand, which essentially can lead to an on-line learning and classification algorithm. Finally, the decision values can be directly utilized for classification purpose.

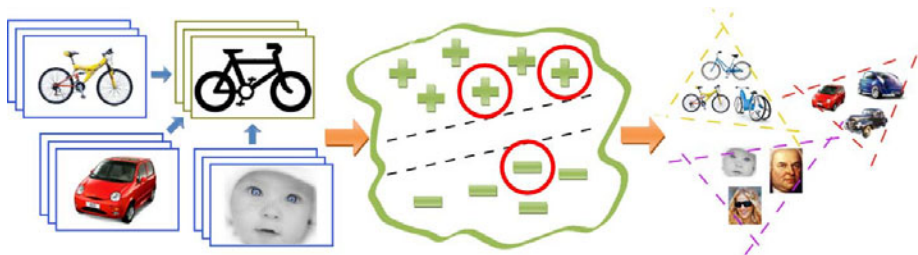


Fig. 1. Illustration of our approach, where yellow rectangles (i.e. bike) denote templates, blue ones (i.e. bikes, cars, faces) denotes training images, a “+” (or “-”) denotes a max-similarity based feature vector generated by a template and a training image with the same category (or different categories). All these features are used to learn sparse features for the template based on a large-margin formulation, and selected features are surrounded by the red circles. Finally, category classifiers, denoted by the triangles, are generated based on the sparse features in the training images.

Considering this, we propose our sparse feature on-line learning approach for image classification as illustrated in Fig. 1. We take each training image as a template, and describe each image as a max-similarity based feature vector in a new feature space defined by the template. Then, a large-margin classifier is trained on-line using linear programming to select sparse features and learn their weights automatically for each template. Eventually, Naive-Bayes (NB) or K Nearest Neighbor (KNN) classifiers are employed to assign each query image to the category with the largest score.

The main contribution of this paper is that our approach can achieve a good balance between system efficiency and classification accuracy.

- **Efficiency:** Several factors improve the efficiency of our approach, i.e. on-line learning of the large-margin classifiers, sparse features and direct usage of margins in classification.
- **Accuracy:** The discriminative power of selected sparse features with their weights has great impact on the classification accuracy of our approach.

The rest of the paper is organized as follows. In Section 2, related work is reviewed. In Section 3, our sparse feature on-line learning algorithm is explained

in detail, including the max-similarity image representation and the learning of large-margin classifiers. In Section 3.3, Naive-Bayes and K Nearest Neighbor category classifiers are described for categorization. In Section 4, our experimental results are shown and compared with others in terms of efficiency and classification accuracy, and finally Section 5 concludes the paper.

2 Related Work

In general, finding feature correspondences between images can be formulated as graph matching problems, where each image is considered as a graph and each feature point is considered as a vertex in the graph. Caetano *et al.* [2] utilized graph learning algorithms to improve the feature correspondence accuracy based on local structures of the graphs. Torresani *et al.* [3] proposed an energy minimization function based on the feature vectors and their local spatial to find a matching sequence. Chen *et al.* [4] defined feature correspondences based on a max-margin formulation in a structured prediction setting to minimize the classification loss. Zhang *et al.* [5] formulated the image matching as bipartition graph matching. These approaches suffer from the high computational complexity, which limits their applications in large-size real image datasets.

A simple way to find feature correspondences is to locate the nearest neighbor for each feature in a template. Recently, several papers suggest that clever usage of the nearest neighbor approaches can improve the classification and detection performance [6, 7, 8, 9, 10]. For instance, Zhang *et al.* [6] proposed an SVM-KNN classifier which first locates K nearest neighbors for a query sample in the feature space and then trains a local multi-class SVM on the set of K neighbors for classification. Boiman *et al.* [7] proposed a Naive-Bayes Nearest-Neighbor (NBNN) classifiers to compute the “Image-to-Class” distances instead of the commonly used “Image-to-Image” distances. Yuan *et al.* [8] utilized the nearest neighbor approaches to accelerate the action detection process. In these works, the nearest neighbors are defined by the distances, while in our approach, we locate the nearest neighbor of each feature by the feature similarity measurement, which allows us to learn the large-margin classifiers directly.

Local Binary Pattern (LBP) was originally introduced by Ojala *et al.* [11] to reflect the intensity relationship between a pixel and its surroundings using 0 and 1. Torresani *et al.* [12] extended this idea to build Boolean features consisting of disjunctions of conjunctions (“OR”s of “AND”s) for scalable image retrieval. These papers demonstrate the discriminative power and efficient learning of binary features by comparing the data with “templates” (*e.g.* In LBP, the centers of local patches will be the templates.). Our large-margin classifiers borrow the similar idea, but return real numbers (margins) instead of binaries to show how likely a query image is in the same category as the templates.

Most related work to ours is the local distance function learning approaches proposed by Frome *et al.* [9, 10]. In [9], a focal learning of local distance function approach tended to maximize the margins amongst the image *triplets*, each of

whom contains one template and two queries, to learn the non-negative weights for features in the templates. However, it suffers from the following drawbacks: 1. It costs huge amount of memory (or disk space); 2. The query-template distances based on the learned weights cannot be used directly for classification purpose. To overcome 2, in [10] the authors proposed another global learning approach based on a new image triplet representation, where there are two templates and one query, to make sure that the non-negative weight learning process for each template is connected to all the other templates so that the classification can be performed directly using the query-template distances. Notice that the training phase of this approach is off-line.

In contrast, 1. our approach adopts image *pairs* rather than triplets to greatly reduce the memory (or disk space) requirement; 2. our approach tends to learn more sparse features; 3. the weights for sparse features are arbitrary values, which makes our learning more flexible and efficient; 4. our large-margin classifiers can be trained either on-line or off-line; 5. the image classification can be performed directly based on the learned margins.

3 On-Line Learning of Sparse Features

Our approach takes each training image as a template and maps all the other training images into a new feature space defined by the template to learn sparse features. It is an on-line algorithm since the sparse feature learning process of a template is independent of the new-coming training images. The on-line learning property allows our approach to be trained more flexibly and efficiently (see our experimental section) than many off-line learning approaches [10, 13, 14].

3.1 Max-Similarity Based Image Representation

Max-Similarity based image representation tries to capture the similarities between images by finding the feature correspondences between a query image and the template. Each query image is described in a new feature space defined by the template. Unlike [9, 10], the max-similarity based feature vector of each image consists of the maximum similarities between each feature in the template and the features in the query, rather than the minimum distances.

Given a template $T = \{t_1, t_2, \dots, t_m\}$ containing m features and a query image $Q = \{q_1, q_2, \dots, q_n\}$ containing n features, the *max-similarity based feature vector* of Q with respect to T , denoted $f_{Q \rightarrow T}$, is defined as follows.

$$f_{Q \rightarrow T_i} = \max_{j=1, \dots, n} K(q_j, t_i) \quad i = 1, \dots, m \quad (1)$$

where $f_{Q \rightarrow T_i}$ denotes the value at the i^{th} dimension of $f_{Q \rightarrow T}$ and $K(\cdot, \cdot)$ denotes any similarity measurement function, *e.g.* the linear kernel, gaussian kernels, and histogram intersection kernel [14].

3.2 Sparse Feature Learning Algorithm

By selecting sparse discriminative features, our sparse feature learning algorithm is to measure the likelihood of two images belonging to the same category. To obtain sparsity, our algorithm assigns non-zero weights to only a small portion of features in order to accelerate the classification speed while achieving a good performance.

For sparsity learning, the large-margin criterion is quite useful for feature selection, which tries to maximize the margin between the positive data and the negative data. 1-norm Support Vector Machines (SVM) [15] and Linear Programming Boosting (LPBoost) [16] are two efficient learning algorithm based on the large-margin criterion, both of which can be solved using LP. One of their main differences is that in 1-norm SVM the learned weights are real numbers, while in LPBoost the learned weights are non-negative and their summation is equal to 1. In our approach, we adopt 1-norm Support Vector Machines (SVM) [15] because real-number weights give us more chances to learn better large-margin classifiers for image classification.

Mathematically, a binary soft-margin 1-norm SVM can be formulated as follows.

$$\begin{aligned} \min_{\mathbf{w}, \epsilon, b} \quad & \|\mathbf{w}\|_1 + C \sum_{i=1}^S \epsilon_i \\ \text{s.t.} \quad & \forall i, y_i(\mathbf{w}' f_i + b) \geq 1 - \epsilon_i, \epsilon_i \geq 0 \end{aligned} \quad (2)$$

where $\|\cdot\|_1$ and $'$ denote the 1-norm and vector transpose operators, \mathbf{w} and ϵ denote the weight and error vectors, C is the regularization parameter, which is a predefined non-negative constant, y_i denotes the label of the max-similarity based feature vector f_i , S denotes the total number of feature vectors, and b denotes the bias term. The value of y_i is dependent on the category of the query image and the template: $y_i = 1$ if the query image is in the same category, otherwise, $y_i = -1$.

Eqn. 2 can be solved using LP based on its equivalent formulation shown in Eqn. 3.

$$\begin{aligned} \min_{\mathbf{w}, \mathbf{v}, \epsilon, b} \quad & \mathbf{1}' \mathbf{v} + C \sum_{i=1}^S \epsilon_i \\ \text{s.t.} \quad & \forall i, y_i(\mathbf{w}' f_i + b) \geq 1 - \epsilon_i \\ & \mathbf{v} \succeq \mathbf{w} \succeq -\mathbf{v}, \mathbf{v} \succeq 0, \epsilon_i \geq 0 \end{aligned} \quad (3)$$

where $\mathbf{1}$ denotes a vector of ones, \mathbf{v} is a non-negative vector, and \succeq is the element-wise operator of \geq .

In order to reduce the computational complexity, we instead optimize the following problem proposed by Fung and Mangasarian [17] in Eqn. 4 since it

contains fewer constraints than Eqn. 3 and returns its upper bound.

$$\begin{aligned} \min_{\mathbf{w}_1, \mathbf{w}_2, \epsilon, b} \quad & \mathbf{1}'(\mathbf{w}_1 + \mathbf{w}_2) + C \sum_{i=1}^S \epsilon_i \\ \text{s.t.} \quad & \forall i, y_i[(\mathbf{w}_1 - \mathbf{w}_2)' f_i + b] \geq 1 - \epsilon_i \\ & \mathbf{w}_1 \succeq 0, \mathbf{w}_2 \succeq 0, \epsilon_i \geq 0 \end{aligned} \quad (4)$$

where \mathbf{w}_1 and \mathbf{w}_2 are two non-negative weight vectors and the weight vector \mathbf{w} for features is defined as $\mathbf{w} = \mathbf{w}_1 - \mathbf{w}_2$.

In our experiments, we find that when the dimension of max-similarity based feature vectors is large enough, say 500D, the solutions of Eqn. 4 are always the same as those of the original 1-norm SVM, but much faster. Therefore, the sparsity property of 1-norm SVM is still kept when solving Eqn. 4.

Using the learned weights \mathbf{w} , the margin between template T and query Q , $F_{Q \rightarrow T}$, is defined as follows. This indicates how likely T and Q belong to the same category.

$$F_{Q \rightarrow T} = (\mathbf{w}_1 - \mathbf{w}_2)' f_{Q \rightarrow T} + b \quad (5)$$

3.3 Image Classification

For each training image, the sparse feature learning algorithm is performed. Using Eqn. 4, we can learn sparse features from the training images, within the same category, and calculate category scores directly for classification. However, we do not advocate to do this because: it has much higher computational complexity and the learning process tends to be more likely over-fitting due to the much higher complexity of the model (much more features), which will lead to poor classification accuracy. We employ a Naive-Bayes or K Nearest Neighbor classifier to perform the classification based on the margins between the query images and the training images.

Given a query image Q , L categories and M_l ($l = 1, 2, \dots, L$) training images $\mathbf{T} = \{T_1, T_2, \dots, T_{M_l}\}$ with their corresponding weights $\mathbf{p}(\mathbf{T}, l)$, the score of a *Naive-Bayes category classifier* (NBCC) for category l , denoted $\mathbb{F}_{Q \rightarrow l}$, is defined as follows.

$$\mathbb{F}_{Q \rightarrow l} = \frac{\sum_{i=1}^{M_l} p(T_i, l) F_{Q \rightarrow T_i}}{\sum_{i=1}^{M_l} p(T_i, l)} \quad (6)$$

In our experiments, since we do not know any prior knowledge about the training data, we adopt the uniform distribution for $\mathbf{p}(\mathbf{T}, l)$. Therefore, Eqn. 6 can be simplified into Eqn. 7 and the category label of Q , l_Q , will be assigned according to Eqn. 8 based on the one-vs-rests criterion.

$$\mathbb{F}_{Q \rightarrow l} = \frac{1}{M_l} \sum_{i=1}^{M_l} F_{Q \rightarrow T_i} \quad (7)$$

$$l_Q = \arg \max_{l=1, \dots, L} \mathbb{F}_{Q \rightarrow l} \quad (8)$$

For a K Nearest Neighbor category classifier (KNNCC), the category score is defined as the number of the training images in the category among the K nearest neighbors of a query. If there are more than one category assignment with the same maximum number among all the categories, Eqn. 7 is able to calculate refiner scores for the category assignments. Eventually the category label of a query is still assigned based on Eqn. 8.

4 Experimental Results

We test our approach for object categorization on Caltech-101 dataset and scene classification on Scene-15 dataset. Each query image is labeled using a one-vs-rests strategy. All features from the images are normalized using l_2 -norm and the max-similarity based feature representation uses the linear kernel due to its computational efficiency. Parameter C in Eqn. 4 is fixed as 10^5 without tuning. Our classification results are the average mean classification accuracy across different categories after several runs. The computational time is calculated on our unoptimized MATLAB implementation on a 2.33GHz Core 2 Duo CPU.

4.1 Caltech-101

Caltech-101 consists of over 9000 images in 101 object categories plus a background category. The numbers of randomly selected training images per category are 1, 5, 10, 15, 20, 25, 30, and the rest are taken as the query images. Parameter K in K Nearest Neighbor category classifiers is fixed as 2.5 times the number of the training images per category. In our experiments, we adopt the Geometric Blur (GB) [18] descriptors [1] at four scales (10, 20, 30, and 40 pixels) to show the inherent property of combining different descriptors in our approach. Each image is converted to gray-scale and resized so that its larger dimension is 300 pixels. Canny edge detector is then employed to detect the edges in images. Four scales (10, 20, 30, and 40 pixels) are defined for GB descriptors so that at most 500 descriptors are generated each scale, totally no more than 2000 descriptors for each image. Sparse feature learning is performed at each scale separately for each training image.

Fig. 2 shows some examples of learned sparse features at scale 10 pixels using 10 training images and $N = 5$ negative samples per category. The green circles denote the features with positive learned weights and the red circles denote the features with negative learned weights. Due to the effect of sparsity learning, the number of the selected features in each image are much fewer than the feature candidates (usually 500). Meanwhile, it is easy to see that these sparse features can capture the commonality within the categories and the difference between the categories, *e.g.* the nose in a cougar face, the arms of a windsor chair. This reflects the discriminative power of the learned sparse features.

¹ We download the MATLAB code from http://www.cs.berkeley.edu/~aberg/demos/gb_demo.tar.gz.

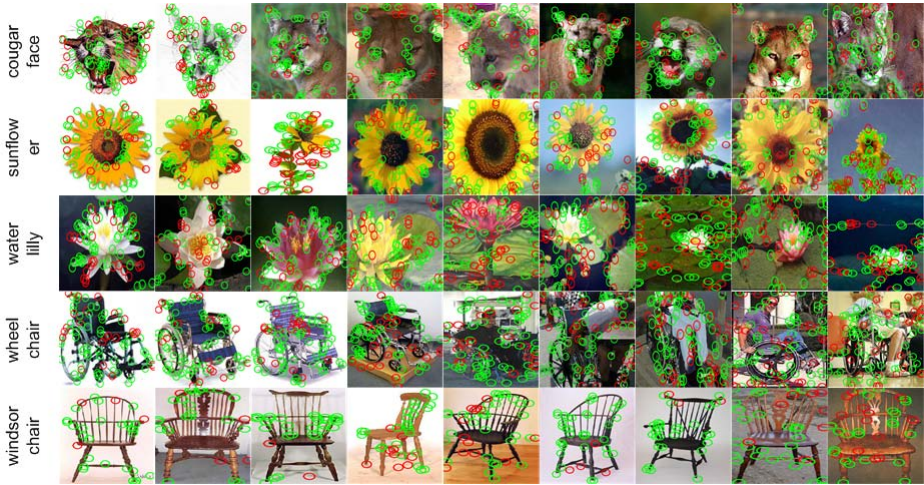


Fig. 2. Some examples of the learned sparse features

Fig. 3 (a) shows the average percentage of selected features using different numbers of training images and Fig. 3 (b) and (c) show our classification accuracy using NBCC and KNNCC. Multi-GB descriptors achieves better classification accuracy than any other GB descriptors. The improvements are about 7.8% for NBCC and 9.9% for KNNCC. Impressively, using 15 training images only for each category and the Multi-GB descriptors, our approach selects $212/1876=11.3\%$ features on average and achieves 53.7% mean accuracy with NBCC and 54.5% with KNNCC. In contrary, [9] achieved 60.3% by selecting 31% features. Using 30 training images per category, the classification accuracy in [19] is 60.1%, while ours are 59.7% using NBCC and 60.4% using KNNCC. We achieved the state-of-the-art on-line learning results on Caltech-101.

The classification speed of our category classifiers is highly dependent on the number of training images and the numbers of their features. With the help

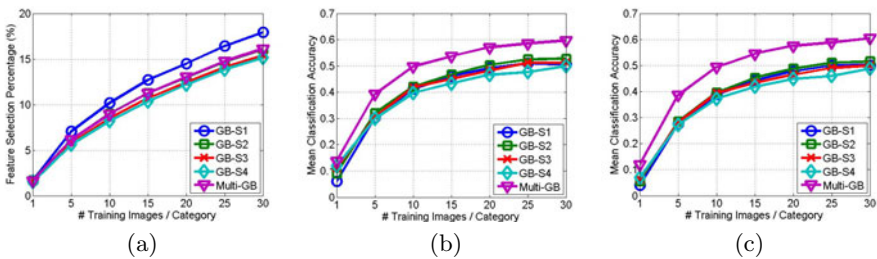


Fig. 3. Our feature selection percentage and mean classification accuracy on Caltech-101. (a) Feature selection percentage, (b) Classification accuracy using NBCC, (c) Classification accuracy using KNNCC.

of sparse feature learning, our classification can be performed very fast. For instance, using 15 training images per category, our approach takes only about 5 *seconds* to classify a query image using NBCC and almost all the computational time is used to calculate max-similarity based feature vectors.

4.2 Scene-15

Scene-15 dataset consists of 4385 images in 15 scene categories and the image resolution is roughly about 250×250 pixels. We first generate 384D OpponentSIFT (OS) descriptors [20] and 100 images are randomly selected as the training data from each category. The number of negative samples per category is set to 30 and parameter K in KNNCC is fixed to 90. Using all features at different scales, our approach selects $486/1731=28.1\%$ features on average for each image and achieves 74.4% using NBCC and 73.8% using KNNCC. We also compare our results with other approaches listed in Table 1. Our classification accuracy is

Table 1. Comparison of classification accuracy for different approaches on Scene-15 dataset (%)

OS-10+NBCC	69.0	OS-10+KNNCC	68.6
OS-20+NBCC	70.5	OS-20+KNNCC	70.1
OS-30+NBCC	68.1	OS-30+KNNCC	68.0
Multi-OS+NBCC	74.4	Multi-OS+KNNCC	73.8
C4+pLSA+SVM [21]	72.6	Co-Clustering [22]	76.4
SPM [14]	74.8	Low-dimensional Feature [23]	72.2

again comparable to the state-of-the-art results.

5 Conclusion

In this paper, an efficient sparse feature on-line learning approach is proposed for image classification. It solves a large-margin formulation using Linear Programming on the max-similarity based image representation. The learned margins indicate how likely the query images and the training images are in the same category. They can be directly involved in a Naive-Bayes or K Nearest Neighbor category classifier to perform the classification. A good balance between efficiency and classification accuracy is the most important characteristic of our approach. This has been demonstrated in our experiments on the Caltech-101 and Scene-15 datasets. Our approach learns a small portion of discriminative features to achieve comparable classification accuracy with the state-of-the-art results with efficient learning and classification algorithms, and it can be applied automatically either on-line or off-line.

References

1. Rosch, E.: Natural categories. *Cognitive Psychology* 4, 328–350 (1973)
2. Caetano, T., Cheng, L., Le, Q., Smola, A.: Learning graph matching. In: *ICCV 2007* (2007)
3. Torresani, L., Kolmogorov, V., Rother, C.: Feature correspondence via graph matching: Models and global optimization. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008*, Part II. LNCS, vol. 5303, pp. 596–609. Springer, Heidelberg (2008)
4. Chen, L., McAuley, J., Feris, R., Caetano, T., Turk, M.: Shape classification through structured learning of matching measures. In: *CVPR 2009*, pp. 365–372 (2009)
5. Zhang, Z., Li, Z.N., Drew, M.S.: Learning image similarities via probabilistic feature matching. In: *ICIP*, pp. 1857–1860 (2010)
6. Zhang, H., Berg, A., Maire, M., Malik, J.: Svm-knn: Discriminative nearest neighbor classification for visual category recognition. In: *CVPR 2006*, vol. II, pp. 2126–2136 (2006)
7. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: *CVPR 2008*, pp. 1–8 (2008)
8. Yuan, J., Liu, Z., Wu, Y.: Discriminative subvolume search for efficient action detection. In: *CVPR 2009*, pp. 2442–2449 (2009)
9. Frome, A., Singer, Y., Malik, J.: Image retrieval and classification using local distance functions. In: *NIPS 2006*, pp. 417–424. MIT Press, Cambridge (2006)
10. Frome, A., Singer, Y., Sha, F., Malik, J.: Learning globally-consistent local distance functions for shape-based image retrieval and classification. In: *ICCV 2007*, pp. 1–8 (2007)
11. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 24(7), 971–987 (2002)
12. Torresani, L., Szummer, M., Fitzgibbon, A.: Learning query-dependent prefilters for scalable image retrieval. In: *CVPR 2009*, pp. 2615–2622 (2009)
13. Gehler, P.V., Nowozin, S.: On feature combination for multiclass object classification. In: *ICCV 2009*, pp. 1–8 (2009)
14. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: spatial pyramid matching for recognizing natural scene categories. In: *CVPR 2006* (2006)
15. Bi, J., Chen, Y., Wang, J.: A sparse support vector machine approach to region-based image categorization. In: *CVPR 2005*, pp. 1121–1128 (2005)
16. Gehler, P., Nowozin, S.: On feature combination for multiclass object classification. In: *ICCV 2009*, pp. 1–8 (2009)
17. Fung, G.M., Mangasarian, O.L.: A feature selection newton method for support vector machine classification. *Comput. Optim. Appl.* 28(2), 185–202 (2004)
18. Berg, A., Malik, J.: Geometric blur and template matching. In: *CVPR 2001*, pp. 607–614 (2001)
19. Ramanan, D., Baker, S.: Local distance functions: A taxonomy, new algorithms, and an evaluation. In: *ICCV 2009* (2009)
20. van de Sande, K.E.A., Gevers, T., Snoek, C.G.M.: Evaluating color descriptors for object and scene recognition. *PAMI* (2010) (in press)
21. Bosch, A., Zisserman, A., Munoz, X.: Scene classification using a hybrid generative/discriminative approach. *PAMI* 30(4), 712–727 (2008)
22. Liu, J., Shah, M.: Scene modeling using co-clustering. In: *ICCV 2007*, pp. 1–7 (2007)
23. Rasiwasia, N., Vasconcelos, N.: Scene classification with low-dimensional semantic spaces and weak supervision. In: *CVPR 2008*, pp. 1–6 (2008)

Classifying Data Considering Pairs of Patients in a Relational Space

Siti Mariam Shafie^{1,2} and Maria Petrou¹

¹ Dept. of Electrical & Electronic Engineering, Imperial College London, South Kensington Campus, London SW7 2AZ, UK

² Dept. of Computer and Communication Systems Engineering, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia
{sitimariam.shafie07,maria.petrou}@imperial.ac.uk

Abstract. In this paper, we demonstrate the use of relational space to classify microarray gene expression data. We also show that the transformation of real valued data to binary data is able to produce better class separation with fewer genes.

Keywords: gene selection, classification, relational space.

1 Introduction

Microarray data has become widely used in classification or clustering of tissues or genes based on gene expression profiles [1][2][3][4]. However, it is difficult to analyze the data because they are massive. Therefore, computational tools have become very important in analysing such data, especially in reducing the number of features or genes [5][6]. Several methods can be used to select a subset of the most important genes. Generally, in the process of selection, the genes are ranked on the basis of scores, correlation coefficient, mutual information and sensitivity analysis. Such analyses usually use real valued data. In this work, we focus on the analysis of gene expression in the binary domain considering data pairs. We first transform the real valued data into binary data and then classify the data in a relational space. The microarray data used are for breast cancer [1]. However, we demonstrate the use of relational spaces also for leukemia [2] and colon cancer [3] data sets. In each case the problem is to separate the microarray data into two classes.

2 Methodology

2.1 Gene Selection

Let n be the number of genes, m^A be the number of patients in group A and m^B be the number of patients in group B. Let g_i be the expression values of gene i where $i = 1, 2, \dots, n$. Then the profile of gene i in group A is

$$G_i^A = \left\{ g_i^{P_1}, g_i^{P_2}, g_i^{P_3}, \dots, g_i^{P_{m^A}} \right\} \quad (1)$$

where P_{m^A} is the last person which belongs to group A. Similarly, we define G_i^B as the profile of gene i in group B. To reduce the complexity of the gene expression analysis, we binarise the data, so the positive values become 1 and the negative 0.

A *significant* gene i is defined as a gene that has an average value for at least one of the groups, that is as distinct as possible from μ_i , the average value over both groups. We define for each gene:

$$\Sigma_i^* = \max \{ |\mu_i^A - \mu_i|, |\mu_i^B - \mu_i| \} \quad (2)$$

Based on this value, we can set a threshold that enables us to select the genes that have the largest gap in their values from the rest.

2.2 Creation of Binary Codes

We can encode each sequence of genes with a binary number. For example, let us consider patient $P_{A,1}$ in group A who has

$$P_{A,1} = [100100]. \quad (3)$$

Then, this patient can be represented by the binary number 100100, which in the decimal system is equal to 36. With different order of the significant genes, we can create different such numbers representing the same patient. The order of the genes is not immunologically important [7]. This enables us to code the patient in several ways. Note that each ordering of the genes is equivalent to giving different weights to the different genes. We can create several combinations of features based on different codes. We may try to find the best combination of features that gives the minimum possible classification error. To do this, we calculate the frequencies of occurrence of each of the selected genes across the patients in each group and in all patients. The genes are ordered from the least frequent (L.F.) to the most frequent (M.F.) in each row. Then we may choose convenient permutations of the genes that allow us to distinguish groups A and B. We call these permutation codes, and distinguish them with letters from the Greek alphabet.

2.3 The Relational Space

The relational space [7] is a space where pairs of patients are represented as opposed to individual patients. The relational space has two axes: along the first one we measure the one code of one patient and along the other another code of its paired patient, who is from the same group. Table 1 shows an example of three patients in group A with their α and β codes. These numbers form two sequences: the α - *sequence* and the β - *sequence* of group A. Both sequences are then sorted in increasing order as shown in Table 2. From these two sorted sequences, we form the pairs of patients (P_2, P_1) , (P_1, P_3) and (P_3, P_2) .

Therefore, using two different sequences, we can plot the pairs of the training data in 2D spaces. Fig. 1 shows an example relational space from the breast

Table 1. Representing the data in binary codes

Patient	Binary data	α - code	β - code
P_1	100100	$(100100)_2 = 36_{10}$	$(000011)_2 = 3_{10}$
P_2	100001	$(0000101)_2 = 5_{10}$	$(100010)_2 = 34_{10}$
P_3	010100	$(110000)_2 = 48_{10}$	$(000100)_2 = 5_{10}$

Table 2. Sorted binary codes

Patient	sorted α - sequence	Patient	sorted β - sequence
P_2^α	$(0000101)_2 = 5_{10}$	P_1^β	$(000011)_2 = 3_{10}$
P_1^α	$(100100)_2 = 36_{10}$	P_3^β	$(000100)_2 = 5_{10}$
P_3^α	$(110000)_2 = 48_{10}$	P_2^β	$(100010)_2 = 34_{10}$

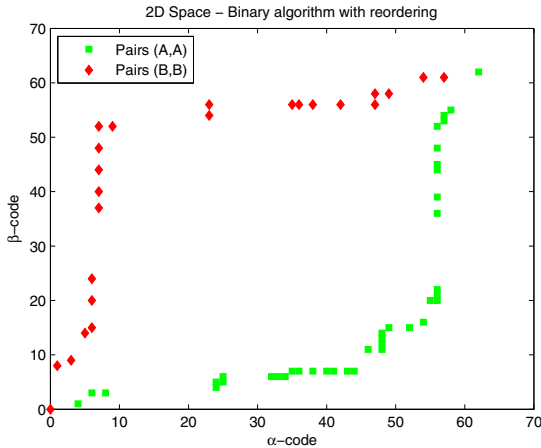


Fig. 1. Relational space obtained with sorted α - code and sorted β - code of pairs of patients

cancer data. With the pairwise structure, surprisingly, we can see clearly that the two groups are separable. In order to show the significance of the pairs we create, we plot randomly paired patients, and each patient paired with herself in Fig. 2 and 3, respectively. We notice that when the order is ignored, the pairs of the two groups are not separable. This suggests that the sorting part in the process of pair creation plays an important role in the construction of a separable relational space.

2.4 Classifier

Let us assume that we have to classify a new patient. We first compute the codes of this patient. Then we add this patient to each group of training patients, in

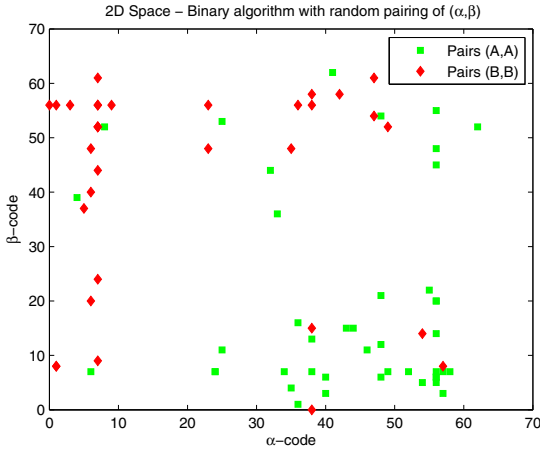


Fig. 2. Random pairing of patients in the relational space of α and β codes

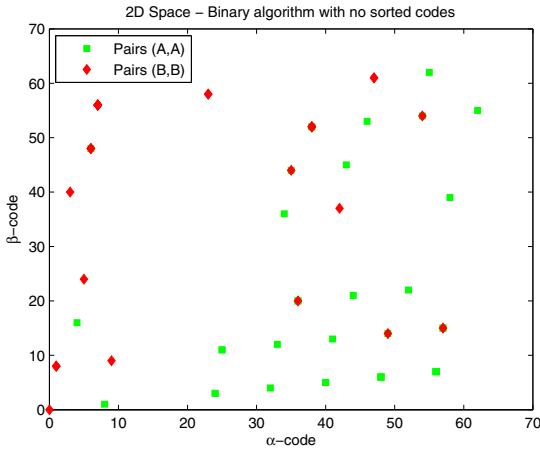


Fig. 3. Pairing each person with herself in the relational space of α and β codes

turn. As the new patient is added, the ordered sequence of codes of the patients is disturbed. When we then pair the patients again, with the new patient included, we shall have a certain number of new pairs created, while other pairs remain identical, as they were before the insertion of the new patient. We may then try to classify the new pairs created, using the k-nearest neighbour classifier in the relational space, and check what fraction of the new pairs created are correctly classified. We shall classify the new patient to the least disturbed group, i.e. the group for which most of the new pairs are correctly classified.

3 Implementation And Experimental Results

The implementation and results are described in detail for the breast cancer data set while for leukemia and colon cancer data set, we only present the classification results briefly.

3.1 Breast Cancer Data Set

The data have been produced by van't Veer et al. [1] where 43 patients diagnosed with breast cancer survived for more than 5 years (group A) and 33 patients died within 5 years (group B). In their study, van't Veer et al. identified 70 genes as a powerful prognostic marker that can be used to classify the patients into the two groups. We analyse these 70 genes here and examine whether we can reduce the number of genes that can be used to classify the patients into the two groups with as small as possible classification error. There are also data of 19 patients as test data. We split the training data into two subsets: training and evaluation. We keep 25 patients from each class as training and the remaining 26 patients for evaluation. The training set will give us the reference points in relation to which new points will be classified, using nearest neighbour classification. The evaluation set will allow us to work out which threshold in the selection of significant genes and which codes produce the best results. Then this threshold and the corresponding codes will be adopted and used to classify the 19 test data. We must note that changing the threshold changes the number of genes selected to represent each patient. So, it is not possible to use the same codes for the same thresholds. In any case, we use the same procedure we described in section 2.2 to create codes for the different thresholds. To distinguish the different codes created for different thresholds, we give a suffix to each Greek letter that represents a code, the same as the threshold used. So, for example, we have codes $\alpha_{0.22}$, $\alpha_{0.23}$ and $\alpha_{0.24}$ to represent codes for threshold 0.22, 0.23 and 0.24 respectively, retaining 9, 6 and 5 genes, respectively.

Table 3. Codes that are used in these experiments. Codes corresponding to different thresholds used to identify the significant genes have different suffixes.

Code	Genes used
$\alpha_{0.22}$	{39,15,6,42,1,69,7,27,65}
$\beta_{0.22}$	{15,6,39,42,27,1,69,7,65}
$\gamma_{0.22}$	{6,39,15,42,7,27,1,69,65}
$\alpha_{0.23}$	{6,39,15,65,42,27}
$\beta_{0.23}$	{42,65,27,6,39,15}
$\gamma_{0.23}$	{42,6,39,15,27,65}
$\delta_{0.23}$	{6,42,15,39,27,65}
$\xi_{0.23}$	{39,6,15,27,42,65}
$\alpha_{0.24}$	{27,42,15,39,6}
$\beta_{0.24}$	{42,27,6,15,39}
$\gamma_{0.24}$	{6,42,27,39,15}

Table 4. Classification result of the evaluation set in various relational spaces

Relational space	Accuracy
$(\alpha_{0.22}, \beta_{0.22})$	76.92%
$(\alpha_{0.22}, \gamma_{0.22})$	88.46%
$(\beta_{0.22}, \gamma_{0.22})$	92.31%
$(\alpha_{0.23}, \beta_{0.23})$	88.46%
$(\gamma_{0.23}, \delta_{0.23})$	88.46%
$(\gamma_{0.23}, \xi_{0.23})$	88.46%
$(\alpha_{0.24}, \beta_{0.24})$	69.23%
$(\alpha_{0.24}, \gamma_{0.24})$	73.08%
$(\beta_{0.24}, \gamma_{0.24})$	76.92%

Table 5. Confusion matrix of classifying the 19 testing data in relational space $(\beta_{0.22}, \gamma_{0.22})$

	A	B
True A	6	1
True B	2	10
Accuracy	84.21%	

Table 3 lists the various codes that will be used in these experiments. Table 4 gives the classification results of the evaluation set obtained for various relational spaces. We observe that the best classification rate was obtained for the relational space $(\beta_{0.22}, \gamma_{0.22})$. Using this relational space we then classified the 19 test data. The accuracy was 84.21%. Table 5 is the confusion matrix for this result.

In order to ensure that the use of relational space offers added value to the classification process, we also considered individual codes to create feature spaces. So, we considered classifying the evaluation patients in 1D and 2D feature spaces, where each patient is represented by a single or by 2 codes. We used the same codes, or pairs of codes, we used to create the relational spaces. The results of classifying the evaluation set using nearest neighbour classification are shown in Table 6. We observe that the best results were obtained for $\beta_{0.22}$ feature space. We then used this feature space to classify the test data. The classification accuracy was 68.42%. Table 7 is the confusion matrix for this result. As the best relational space was the $(\beta_{0.22}, \gamma_{0.22})$, we also classified the test data using only the 1D $\gamma_{0.22}$ feature space and the 2D $(\beta_{0.22}, \gamma_{0.22})$ feature space. The accuracy was 57.89% and 68.42%, respectively. The classification accuracy obtained for the same patients and the same number of genes using the van't Veer et al. method was 78.95%.

3.2 Leukemia Data Set

The data consist of 72 samples of acute lymphoblastic leukemia (ALL) and acute myeloid leukemia (AML) [2]. Each sample consists of expression values of

Table 6. Classification results of the evaluation set in various feature spaces

Space	Dimension	Accuracy
$\alpha_{0.22}$	1	92.31%
$\beta_{0.22}$	1	96.15%
$\gamma_{0.22}$	1	80.77%
$(\alpha_{0.22}, \beta_{0.22})$	2	92.31%
$(\alpha_{0.22}, \gamma_{0.22})$	2	92.31%
$(\beta_{0.22}, \gamma_{0.22})$	2	92.31%
$\alpha_{0.23}$	1	88.46%
$\beta_{0.23}$	1	84.62%
$\gamma_{0.23}$	1	88.46%
$\delta_{0.23}$	1	92.31%
$\xi_{0.23}$	1	88.46%
$(\alpha_{0.23}, \beta_{0.23})$	2	84.62%
$(\gamma_{0.23}, \delta_{0.23})$	2	92.31%
$(\gamma_{0.23}, \xi_{0.23})$	2	92.31%
$\alpha_{0.24}$	1	52.63%
$\beta_{0.24}$	1	57.89%
$\gamma_{0.24}$	1	47.37%
$(\alpha_{0.24}, \beta_{0.24})$	2	57.89%
$(\alpha_{0.24}, \gamma_{0.24})$	2	52.63%
$(\beta_{0.24}, \gamma_{0.24})$	2	52.63%

Table 7. Confusion matrix of classifying the 19 testing data in feature space $\beta_{0.22}$

	A	B
True A	6	1
True B	5	7
Accuracy 68.42%		

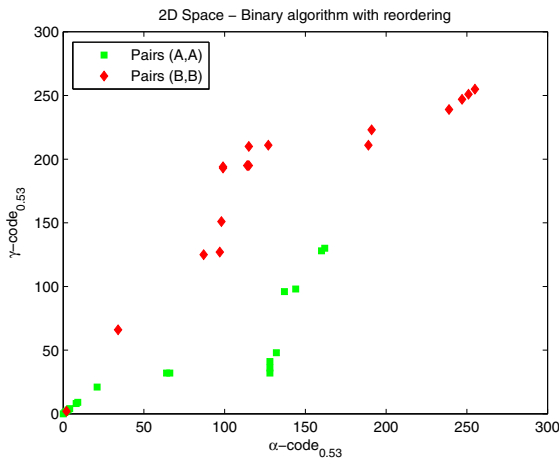


Fig. 4. The relational space for the leukemia data set with 8 selected genes

7129 genes. Golub et al. identified 50 significant genes based on correlation of class distinction. We use these genes in our experiments to examine whether a relational space is able to classify the data. Figure 4 shows the best relational space with 8 selected genes that produce classification accuracy of 91.67% using the leave-one-out method.

3.3 Colon Cancer Data Set

These data were obtained from 22 normal and 40 tumor colon samples [3]. Each sample consists of 2000 genes with the highest minimal intensity across the samples. Each gene is normalized to have mean 0 across the samples and the standard deviation is 1. Figure 5 shows a relational space with 6 selected genes. The classification result using the relational space is 87.1% accuracy using the leave-one-out method.

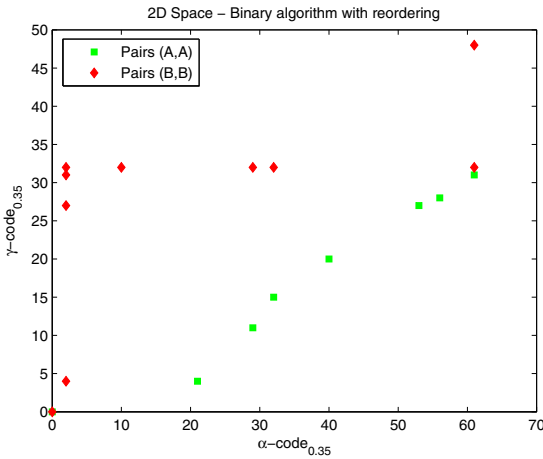


Fig. 5. The relational space for the colon data set with 6 selected genes

4 Conclusions

In this paper, we demonstrated how the use of binary gene expression data and a relational space may result in a representational space where the classes are better separated. We showed that with a small number of genes we can achieve much better performance in the relational space than in the feature space, even when the same data representation can be used. These conclusions were validated using three different data sets, from totally different medical conditions.

References

1. van't Veer, L.J., Dai, H., van de Vijver, M.J., He, Y.D., Kerkhoven, A.M., Roberts, C., Linsley, P.S., Bernards, R., Friend, S.H.: Gene expression profiling predicts clinical outcome of breast cancer. *Nature* 415, 530–536 (2002)
2. Golub, T.R., Slonim, D.K., Tamayo, P., Huard, C., Gaasenbeek, M., Mesirov, J.P., Coller, H., Loh, M.L., Downing, J.R., Caligiuri, M.A., Bloomfield, C.D., Lander, E.S.: Molecular Classification of Cancer: Class Discovery and Class Prediction by Gene Expression Monitoring. *Science* 286, 531–537 (1999)
3. Alon, U., Barkai, N., Notterman, D.A., Gish, K., Ybarra, S., Mack, D., Levine, A.J.: Broad patterns of gene expression revealed by clustering analysis of tumor and normal colon tissues probed by oligonucleotide arrays. *PNAS* 96(12), 6745–6750 (1999)
4. Furey, T.S., Cristianini, N., Duffy, N., Bednarski, D.W., Schummer, M., Haussler, D.: Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics* 16(10), 906–914 (2000)
5. Quackenbush, J.: Computational analysis of microarray data. *Nat. Rev. Genet.* 2, 418–427 (2001)
6. Dudoit, S., Fridlyand, J., Speed, T.P.: Comparison of discrimination methods for the classification of tumors using gene expression data. *Journal of the American Statistical Association* 97, 77–87 (2002)
7. Pintus, P., Petrou, M.: Relational space classification for malaria diagnosis. *Pattern Analysis and Applications*, under review (2008)

Hierarchical Spatial Matching Kernel for Image Categorization

Tam T. Le¹, Yousun Kang², Akihiro Sugimoto²,
Son T. Tran¹, and Thuc D. Nguyen¹

¹ University of Science, VNU-HCMC, Vietnam
{[tttam](mailto:tttam@fit.hcmus.edu.vn),[ttson](mailto:ttson@fit.hcmus.edu.vn),[ndthuc](mailto:ndthuc@fit.hcmus.edu.vn)}@fit.hcmus.edu.vn

² National Institute of Informatics, Tokyo, Japan
{[yskang](mailto:yskang@nii.ac.jp),[sugimoto](mailto:sugimoto@nii.ac.jp)}@nii.ac.jp

Abstract. Spatial pyramid matching (SPM) has been one of important approaches to image categorization. Despite its effectiveness and efficiency, SPM measures the similarity between sub-regions by applying the bag-of-features model, which is limited in its capacity to achieve optimal matching between sets of unordered features. To overcome this limitation, we propose a hierarchical spatial matching kernel (HSMK) that uses a coarse-to-fine model for the sub-regions to obtain better optimal matching approximations. Our proposed kernel can robustly deal with unordered feature sets as well as a variety of cardinalities. In experiments, the results of HSMK outperformed those of SPM and led to state-of-the-art performance on several well-known databases of benchmarks in image categorization, even when using only a single type of feature.

Keywords: kernel method, hierarchical spatial matching kernel, image categorization, coarse-to-fine model.

1 Introduction

Image categorization is the task of classifying a given image into a suitable semantic category. The semantic category can be defined as the depicting of a whole image such as a forest, a mountain or a beach, or of the presence of an interesting object such as an airplane, a chair or a strawberry. Among existing methods for image categorization, the bag-of-features (BoF) model is one of the most popular and efficient. It considers an image as a set of unordered features extracted from local patches. The features are quantized into discrete visual words, with sets of all visual words referred to as a dictionary. A histogram of visual words is then computed to represent an image. One of the main weaknesses in this model is that it discards the spatial information of local features in the image. To overcome it, spatial pyramid matching (SPM) [9], an extension of the BoF model, utilizes the aggregated statistics of the local features on fixed sub-regions. It uses a sequence of grids at different scales to partition the image into sub-regions, and then computes a BoF histogram for each sub-region. Thus,

the representation of the whole image is the concatenation vector of all the histograms.

Empirically, it is realized that to obtain good performances, the BoF model and SPM have to be applied together with specific nonlinear Mercer kernels such as the intersection kernel or χ^2 kernel. This means that a kernel-based discriminative classifier is trained by calculating the similarity between each pair of sets of unordered features in the whole images or in the sub-regions. It is also well known that numerous problems exist in image categorization such as the presence of heavy clutter, occlusion, different viewpoints, and intra-class variety. In addition, the sets of features have various cardinalities and are lacking in the concept of spatial order. SPM embeds a part of the spatial information over the whole image by partitioning an image into a sequence of sub-regions, but in order to measure the optimal matching between corresponding sub-regions, it still applies the BoF model, which is known to be confined when dealing with sets of unordered features.

In this paper, we propose a new kernel function based on the coarse-to-fine approach and we call it a hierarchical spatial matching kernel (HSMK). HSMK allows not only capturing spatial order of local features, but also accurately measuring the similarity between sets of unordered local features in sub-regions. In HSMK, a coarse-to-fine model on sub-regions is realized by using multi-resolutions, and thus our feature descriptors capture not only the local details from fine resolution sub-regions, but also global information from coarse resolution ones. In addition, matching based on our coarse-to-fine model involves a hierarchical process. This indicates that a feature that does not find its correspondence in a fine resolution still has a possibility of having its correspondence in a coarse resolution. Accordingly, our proposed kernel can achieve a better optimal matching approximation between sub-regions than SPM.

2 Related Work

Many recent methods have been proposed to improve the traditional BoF model. Generative methods [12] model the co-occurrence of visual words while discriminative visual words learnings [13,20] or sparse coding methods [11,19] improve the dictionary in terms of discriminative ability or lower reconstruction error instead of using the quantization by K-means clustering. On the other hand, SPM captures the spatial layout of features ignored in the BoF model. Among these improvements, SPM is particularly effective as well as being easy and simple to construct. It is utilized as a major part in many state-of-the-art frameworks in image categorization [3].

SPM is often applied with a nonlinear kernel such as the intersection kernel or χ^2 kernel. This requires high computation and large storage. Maji *et al.* [12] proposed an approximation to improve efficiency in building the histogram intersection kernel, but efficiency can be attained merely by using pre-computed auxiliary tables which are considered as a type of pre-trained nonlinear support vector machines (SVM). To give SPM the linearity needed to deal with large

datasets, Yang [19] proposed a linear SPM with sparse coding (ScSPM), in which a linear kernel is chosen instead of a nonlinear kernel due to the more linearly separable property of sparse features. Wang & Wang [18] proposed a multiple scale learning (MSL) framework in which multiple kernel learning (MKL) is employed to learn the optimal weights instead of using predefined weights of SPM.

Our proposed kernel concentrates on improvement of the similarity measurement between sub-regions by using a coarse-to-fine model instead of the BoF model used in SPM. We consider the sub-regions on a sequence of different resolutions as the pyramid matching kernel (PMK) [4]. Furthermore, instead of using the pre-defined weight vector for basic intersection kernels to penalize across different resolutions, we reformulate the problem into a uniform MKL to obtain it more effectively. In addition, our proposed kernel can deal with different cardinalities of sets of unordered features by applying the square root diagonal normalization [17] for each intersection kernel, which is not considered in PMK.

3 Hierarchical Spatial Matching Kernel

In this section, we first describe the original formulation of SPM and then introduce our proposed HSMK, which uses a coarse-to-fine model as a basic for improving SPM.

3.1 Spatial Pyramid Matching

Each image is represented by a set of vectors in the D -dimensional feature space. Features are quantized into discrete types called visual words by using K -means clustering or sparse coding. The matching between features turns into a comparison between discrete corresponding types. This means that they are matched if they are in the same type and unmatched otherwise.

SPM constructs a sequence of different scales with $l = 0, 1, 2, \dots, L$ on an image. In each scale, it partitions the image into $2^l \times 2^l$ sub-regions and applies the BoF model to measure the similarity between sub-regions. Let X and Y be two sets of vectors in the D -dimensional feature space. The similarity between two sets at scale l is the sum of the similarity between all corresponding sub-regions:

$$\mathcal{K}_l(X, Y) = \sum_{i=1}^{2^{2l}} \mathcal{I}(X_i^l, Y_i^l), \quad (1)$$

where X_i^l is the set of feature descriptors in the i^{th} sub-region at scale l of the image vector set X . The intersection kernel \mathcal{I} between X_i^l and Y_i^l is formulated as:

$$\mathcal{I}(X_i^l, Y_i^l) = \sum_{j=1}^V \min(\mathcal{H}_{X_i^l}(j), \mathcal{H}_{Y_i^l}(j)), \quad (2)$$

where V is the total number of visual words and $\mathcal{H}_\alpha(j)$ is the number of occurrences of the j^{th} visual word which is obtained by quantizing feature descriptors in the set α . Finally, the SPM kernel (SPMK) is the sum of weighted

similarity over the scale sequence:

$$\mathcal{K}(X, Y) = \frac{1}{2^L} \mathcal{K}_0(X, Y) + \sum_{l=1}^L \frac{1}{2^{L-l+1}} \mathcal{K}_l(X, Y). \quad (3)$$

The weight $\frac{1}{2^{L-l+1}}$ associated with scale l is inversely proportional to the sub-region width at that scale. This weight is utilized to penalize the matching since it is easier to find the matches in the larger regions. We remark that all the matches found at scale l are also included in a finer scale $l - \zeta$ with $\zeta > 0$.

3.2 The Proposed Kernel: Hierarchical Spatial Matching Kernel

To improve efficiency in achieving the similarity measurement between sub-regions, we utilize a coarse-to-fine model on sub-regions by mapping them into a sequence of different resolutions $2^{-r} \times 2^{-r}$ with $r = 0, 1, 2, \dots, R$ as in [4].

X_i^l and Y_i^l are the sets of feature descriptors in the i^{th} sub-regions at scale l of image vector sets X, Y respectively. At each resolution r , we apply the normalized intersection kernel \mathcal{F}^r using the square root diagonal normalization method to measure the similarity as follows:

$$\mathcal{F}^r(X_i^l, Y_i^l) = \frac{\mathcal{I}(X_i^l(r), Y_i^l(r))}{\sqrt{\mathcal{I}(X_i^l(r), X_i^l(r)) \mathcal{I}(Y_i^l(r), Y_i^l(r))}}, \quad (4)$$

where $X_i^l(r), Y_i^l(r)$ are the sets X_i^l, Y_i^l at the resolution r respectively. Note that the histogram intersection between X and itself is equivalent with its cardinality. Thus, letting $\mathcal{N}_{X_i^l(r)}$ and $\mathcal{N}_{Y_i^l(r)}$ be the cardinality of sets $X_i^l(r)$ and $Y_i^l(r)$, the equation (4) is rewritten as:

$$\mathcal{F}^r(X_i^l, Y_i^l) = \frac{\mathcal{I}(X_i^l(r), Y_i^l(r))}{\sqrt{\mathcal{N}_{X_i^l(r)} \mathcal{N}_{Y_i^l(r)}}}. \quad (5)$$

The square root diagonal normalization of the intersection kernel not only satisfies Mercer's conditions [17], but also penalizes the difference in cardinality between sets as in equation (5).

To obtain the synthetic similarity measurement of the coarse-to-fine model, we define the linear combination over a sequence of local kernels, each term of which is calculated using equation (5) at each resolution. Accordingly, the kernel function \mathcal{F} between two sets X_i^l and Y_i^l in the coarse-to-fine model is formulated as:

$$\mathcal{F}(X_i^l, Y_i^l) = \sum_{r=0}^R \theta_r \mathcal{F}^r(X_i^l, Y_i^l) \quad (6)$$

$$\text{where } \sum_{r=0}^R \theta_r = 1, \theta_r \geq 0, \forall r = 0, 1, 2, \dots, R.$$

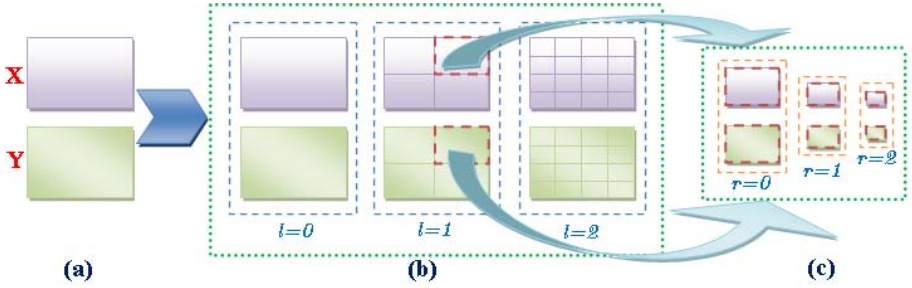


Fig. 1. An illustration for HSMK applied to images X and Y with $L = 2$ and $R = 2$ (a). HSMK first partitions the images into $2^l \times 2^l$ sub-regions with $l = 0, 1, 2$ as SPMK (b). However, HSMK applies the coarse-to-fine model for each sub-region by considering it on a sequence of different resolutions $2^{-r} \times 2^{-r}$ with $r = 0, 1, 2$ (c). Equation (8) with the weight vector achieved from the uniform MKL is applied to obtain better optimal matching approximation between sub-regions instead of using the BoW model as in SPMK.

Moreover, when the linear combination of local kernels is integrated with SVM, it can be reformulated as a MKL problem where basic local kernels are defined as equation (5) across the resolutions of the sub-region as:

$$\begin{aligned}
 \min_{\mathbf{w}_\alpha, w_0, \boldsymbol{\xi}, \boldsymbol{\theta}} \quad & \frac{1}{2} \left(\sum_{\alpha=1}^{\mathfrak{N}} \theta_\alpha \|\mathbf{w}_\alpha\|_2 \right)^2 + \mathcal{C} \sum_{i=1}^N \xi_i \\
 \text{s.t.} \quad & y_i \left(\sum_{\alpha=1}^{\mathfrak{N}} \theta_\alpha \langle \mathbf{w}_\alpha, \Phi_\alpha(\mathbf{x}_i) \rangle + w_0 \right) \geq 1 - \xi_i \\
 & \sum_{\alpha=1}^{\mathfrak{N}} \theta_\alpha = 1, \boldsymbol{\theta} \geq \mathbf{0}, \boldsymbol{\xi} \geq \mathbf{0},
 \end{aligned} \tag{7}$$

where \mathbf{x}_i is an image sample, y_i is the category label for \mathbf{x}_i , N is the number of training samples, $(\mathbf{w}_\alpha, w_0, \boldsymbol{\xi})$ are parameters of SVM, \mathcal{C} is a soft margin parameter defined by users to penalize training errors in SVM, $\boldsymbol{\theta}$ is a weight vector for basic local kernels, \mathfrak{N} is the number of the basic local kernels of the sub-region over the sequence of resolutions, $\boldsymbol{\theta} \geq \mathbf{0}$ means that any entry of vector $\boldsymbol{\theta}$ is nonnegative, $\Phi(\mathbf{x})$ is the function that maps the vector \mathbf{x} into the reproducing Hilbert space and $\langle \cdot, \cdot \rangle$ denotes the inner product. MKL solves the parameters of SVM and the weight vector for basic local kernels simultaneously.

These basic local kernels are analogously defined across resolutions of the sub-region. Therefore, the redundant information between them is high. The experiments in Gehler and Nowozin [3] and especially Kloft *et al.* [7] have shown that the uniform MKL, which is an approximation of MKL into traditional non-linear kernel SVM, is the most efficient for this case in terms of both performance and complexity. Thus, formula (6) with linear combination coefficients obtained from the uniform MKL method becomes:

$$\mathcal{F}(X_i^l, Y_i^l) = \frac{1}{R+1} \sum_{r=0}^R \mathcal{F}^r(X_i^l, Y_i^l). \tag{8}$$

Figure 1 illustrates an application of HSMK with $L = 2$ and $R = 2$. HSMK also maps the sub-regions into a sequence of different resolutions for PMK to obtain better measurement of similarity between them. However, the weight vector is achieved from the uniform MKL. Thus, it is more efficient and theoretical than predefined one in PMK. Furthermore, applying the square root diagonal normalization allows it to robustly deal with differences in cardinality that are not considered in PMK. HSMK is formulated based on SPM in the coarse-to-fine model, which is efficient with sets of unordered feature descriptors, even in the presence of differences in cardinality. Mathematically, the formulation of HSMK is as follows:

$$\begin{aligned} \mathcal{K}(X, Y) &= \frac{1}{2^L} \mathcal{F}_0(X, Y) + \sum_{l=1}^L \frac{1}{2^{L-l+1}} \mathcal{F}_l(X, Y) \\ \text{with } \mathcal{F}_l(X, Y) &= \sum_{i=1}^{2^{2l}} \mathcal{F}(X_i^l, Y_i^l) = \frac{1}{R+1} \sum_{i=1}^{2^{2l}} \sum_{r=0}^R \mathcal{F}^r(X_i^l, Y_i^l). \end{aligned} \quad (9)$$

Briefly, HSMK utilizes the kd -tree algorithm to map each feature descriptor into a discrete visual word, and then the normalized intersection kernel by the square root diagonal method is applied to the histogram of V bins to measure the similarity. We have \mathcal{N} feature descriptors in the D -dimension space, and the kd -tree algorithm costs $O(\log V)$ steps to map feature descriptors. Therefore, the complexity of HSMK is $O(DM \log V)$ with $M = \max(\mathcal{N}_X, \mathcal{N}_Y)$. We note that the complexity of the optimal matching kernel [8] is $O(DM^3)$.

4 Experimental Results

Most recent approaches use local invariant features as an effective means of representing images, because they can well describe and match instances of objects or scenes under a wide variety of viewpoints, illuminations, or even background clutter. Among them, SIFT [10] is one of the most robust and efficient features. To achieve better discriminative ability, we utilize the dense SIFT by operating a SIFT descriptor of 16×16 patches computed over each pixel of an image instead of key points [10] or a grid of points [9]. In addition, to improve robustness, we convert images into gray scale ones before computing the dense SIFT. Dense features have the capability of capturing uniform regions such as sky, water or grass where key points usually do not exist. Moreover, the combination of dense features and the coarse-to-fine model allows images to be represented more exactly since feature descriptors achieves more neighbor information across many levels in resolution. We performed unsupervised K-means clustering on a random subset of SIFT descriptors to build visual words. Typically, we used two different dictionary sizes M in our experiment: $M = 400$ and $M = 800$.

We conducted experiments for two types of image categorization: object categorization and scene categorization. For object categorization, we used the Oxford Flower dataset [14]. To show the efficiency and scalability of our proposed kernel, we also used the large scale object datasets such as CALTECH-101 [2]

and CALTECH-256 [5]. For scene categorization, we evaluated the proposed kernel on the MIT scene [16] and UIUC scene [9] datasets.

4.1 Object Categorization

Oxford Flowers dataset: This dataset contains 17 classes of common flowers in the United Kingdom, collected by Nilsback *et al.* [14]. Each class has 80 images with large scale, pose and light variations. Moreover, intra-class flowers such as irises, fritillaries and pansies are also widely diverse in their colors and shapes. There are some cases of close similarity between flowers of different classes such as that between dandelion and Colts’Foot. In our experiments, we followed the set-up of Gehler and Nowozin [3], randomly choosing 40 samples from each class for training and using the rest for testing. Note that we did not use a validation set as in [14,15] for choosing the optimal parameters. Table 1 shows that our proposed kernel achieved a state-of-the-art results when using a single feature. It outperformed not only SIFT-Internal [15], the best feature for this dataset computed on a segmented image, but also the same feature on SPMK with the optimal weights by MSL [18]. In addition, Table 2 shows that the performance of HSMK also outperformed that of SPMK.

Table 1. Classification rate (%) with a single feature comparison on Oxford Flower dataset (with NN that denotes the nearest neighbour algorithm)

Method	Accuracy (%)
HSV (NN) [15]	43.0
SIFT-Internal (NN) [15]	55.1
SIFT-Boundary (NN) [15]	32.0
HOG (NN) [15]	49.6
HSV (SVM) [3]	61.3
SIFT-Internal (SVM) [3]	70.6
SIFT-Boundary (SVM) [3]	59.4
HOG (SVM) [3]	58.5
SIFT (MSL) [18]	65.3
Dense SIFT (HSMK)	72.9

Table 2. Classification rate (%) comparison between SPMK and HSMK on Oxford Flower dataset

Kernel	$M = 400$	$M = 800$
SPMK	68.09%	69.12%
HSMK	71.76%	72.94%

Caltech datasets: To show the efficiency and robustness of HSMK, we also evaluated its performance on large scale object datasets, i.e., the CALTECH-101 and CALTECH-256 datasets. These datasets feature high intra-class variability,

Table 3. Classification rate (%) comparison on CALTECH-101 dataset

	5	10	15	20	25	30
	training	training	training	training	training	training
Grauman & Darrell [4]	34.8%	44%	50.0%	53.5%	55.5%	58.2%
Wang <i>et al.</i> [18]	-	-	61.4%	-	-	-
Lazebnik <i>et al.</i> [9]	-	-	56.4%	-	-	64.6%
Yang <i>et al.</i> [19]	-	-	67.0%	-	-	73.2%
Boimann <i>et al.</i> [1]	56.9%	-	72.8%	-	-	79.1%
Gehler & Nowozin (MKL) [3]	42.1%	55.1%	62.3%	67.1%	70.5%	73.7%
Gehler & Nowozin (LP- β) [3]	54.2%	65.0%	70.4%	73.6%	75.7%	77.8%
Gehler & Nowozin (LP-B) [3]	46.5%	59.7%	66.7%	71.1%	73.8%	77.2%
Our method (HSMK)	50.5%	62.2%	69.0%	72.3%	74.4%	77.3%

Table 4. Classification rate (%) comparison between SPMK and HSMK on CALTECH-101 dataset

	5	10	15	20	25	30
	training	training	training	training	training	training
SPMK ($M = 400$)	48.18%	58.86%	65.34%	69.35%	71.95%	73.46%
HSMK($M=400$)	50.68%	61.97%	67.91%	71.35%	73.92%	75.59%
SPMK ($M = 800$)	48.11%	59.70%	66.84%	69.98%	72.62%	75.13%
HSMK($M=800$)	50.48%	62.17%	68.95%	72.32%	74.36%	77.33%

poses, and viewpoints. On CALTECH-101, we carried out experiments with 5, 10, 15, 20, 25, and 30 training samples for each class, including the background class, and used up to 50 samples per class for testing. Table 3 compares the classification rate results of our approach with other ones. As shown, our approach obtained the comparable result with that of state-of-the-art approaches even using only a single feature while others used many types of features and complex learning algorithms such as MKL and linear programming boosting (LP-B) [3]. Table 4 shows that the result of HSMK outperformed that of SPMK in this case as well. It should be noted that when the experiment was conducted without the background class, our approach achieved a classification rate of 78.4% for 30 training samples. This shows that our approach is efficient in spite of its simplicity.

On CALTECH-256, we performed experiments with HSMK using 15 and 30 training samples per class, including the clutter class, and 25 samples of each class for testing. We also re-implemented SPMK [5] but used our dense SIFT to enable a fair comparison of SPMK and HSMK. As shown in Table 5, the HSMK classification rate was about 3 percent higher than that of SPMK.

4.2 Scene Categorization

We also performed experiments using HSMK on the MIT Scene (8 classes) and UIUC Scene (15 classes) dataset. In these datasets, we set $M = 400$ as the

Table 5. Classification rate (%) comparison on CALTECH-256 dataset

Kernel	15 training	30 training
Griffin <i>et al.</i> (SPMK) [5]	28.4%	34.2%
Yang <i>et al.</i> (ScSPM) [19]	27.7%	34.0%
Gehler & Nowozin (MKL) [3]	30.6%	35.6%
SPMK	25.3%	31.3%
Our method (HSMK)	27.2%	34.1%

dictionary size. On the MIT Scene dataset, we randomly chose 100 samples per class for training and 100 other samples per class for testing. As shown in Table 6, the classification rate for HSMK was 2.5 percent higher than that of SPMK. Our approach also outperformed other local feature approaches [6] as well as local feature combinations [6] by more than 10 percent, and was better than the global feature GIST [16], an efficient feature in scene categorization.

Table 6. Classification rate (%) comparison on MIT Scene (8 classes) dataset

Method	Accuracy (%)
GIST [16]	83.7
Local features [6]	77.2
Dense SIFT (SPMK)	85.8
Dense SIFT (HSMK)	88.3

On the UIUC Scene dataset, we followed the experiment setup described in [9]. We randomly chose 100 training samples per class and the rest were used for testing. As shown in Table 7, the result of our proposed kernel also outperformed that of SPMK [9] as well as SPM based on sparse coding [19] for this dataset.

Table 7. Classification rate (%) comparison on UIUC Scene (15 classes) dataset

Method	Accuracy (%)
Lazebnik <i>et al.</i> (SPMK) [9]	81.4
Yang <i>et al.</i> (ScSPM) [19]	80.3
SPMK	79.9
Our method (HSMK)	82.2

5 Conclusion

In this paper, we proposed an efficient and robust kernel that we call the hierarchical spatial matching kernel (HSMK). It uses a coarse-to-fine model for sub-regions to improve spatial pyramid matching kernel (SPMK) and thus obtains more neighbor information through a sequence of different resolutions. In

addition, the kernel efficiently and robustly handles sets of unordered features as SPMK and pyramid matching kernel as well as sets having different cardinalities.

Combining the proposed kernel with a dense feature approach was found to be sufficiently effective and efficient. It enabled us to obtain at least comparable results with those by existing methods for many kinds of datasets. Moreover, our approach is simple since it is based on only a single feature with nonlinear support vector machines, in contrast to other more complicated recent approaches based on multiple kernel learning or feature combinations.

In most well-known datasets of object and scene categorization, the proposed kernel was also found to outperform SPMK which is an important component such as a basic kernel in multiple kernel learning. This means that we can replace SPMK with HSMK to improve the performance of frameworks based on basic kernels.

Acknowledgements

This work was performed under the National Institute of Informatics international internship program and also the framework of Memorandum of Understanding between the Vietnam National University of Ho Chi Minh City and the National Institute of Informatics, Japan. This work was in part supported by JST, CREST.

References

1. Boiman, O., Shechtman, E., Irani, M.: In defense of nearest-neighbor based image classification. In: CVPR (2008)
2. Fei-Fei, L., Fergus, R., Perona, P.: Learning generative visual models from few training examples: an incremental Bayesian approach tested on 101 object categories. In: Workshop on Generative-Model Based Vision (2004)
3. Gehler, P., Nowozin, S.: On feature combination for multiclass object classification. In: ICCV, pp. 221–228 (2009)
4. Grauman, K., Darrell, T.: The pyramid match kernel: discriminative classification with sets of image features. In: ICCV, vol. 2, pp. 1458–1465 (2005)
5. Griffin, G., Holub, A., Perona, P.: Caltech-256 object category dataset. Tech. Rep. 7694, California Institute of Technology (2007)
6. Johnson, M.: Semantic Segmentation and Image Search. Ph.D. thesis, University of Cambridge (2008)
7. Kloft, M., Brefeld, U., Laskov, P., Sonnenburg, S.: Non-sparse multiple kernel learning. In: NIPS Workshop on Kernel Learning: Automatic Selection of Kernels (2008)
8. Kondor, R.I., Jebara, T.: A kernel between sets of vectors. In: ICML, pp. 361–368 (2003)
9. Lazebnik, S., Schmid, C., Ponce, J.: Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In: CVPR, vol. 2, pp. 2169–2178 (2006)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2), 91–110 (2004)
11. Mairal, J., Bach, F., Ponce, J., Sapiro, G.: Online dictionary learning for sparse coding. In: ICML, pp. 689–696 (2009)

12. Maji, S., Berg, A., Malik, J.: Classification using intersection kernel support vector machines is efficient. In: CVPR, pp. 1–8 (2008)
13. Moosmann, F., Triggs, B., Jurie, F.: Randomized clustering forests for building fast and discriminative visual vocabularies. In: NIPS Workshop on Kernel Learning: Automatic Selection of Kernels (2008)
14. Nilsback, M.E., Zisserman, A.: A visual vocabulary for flower classification. In: CVPR, vol. 2, pp. 1447–1454 (2006)
15. Nilsback, M.E., Zisserman, A.: Automated flower classification over a large number of classes. In: ICVGIP (2008)
16. Oliva, A., Torralba, A.: Modeling the shape of the scene: A holistic representation of the spatial envelope. IJCV 42, 145–175 (2001)
17. Scholkopf, B., Smola, A.J.: Learning with Kernels: Support Vector Machines, Regularization, Optimization, and Beyond. MIT Press, Cambridge (2001)
18. Wang, S.C., Wang, Y.C.F.: A multi-scale learning framework for visual categorization. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part I. LNCS, vol. 6492, pp. 310–322. Springer, Heidelberg (2011)
19. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: CVPR, pp. 1794–1801 (2009)
20. Yang, L., Jin, R., Sukthankar, R., Jurie, F.: Unifying discriminative visual codebook generation with classifier training for object category recognition. In: CVPR, Los Alamitos, CA, USA, pp. 1–8 (2008)

Feature Selection for Tracker-Less Human Activity Recognition*

Plinio Moreno, Pedro Ribeiro, and José Santos-Victor

Instituto de Sistemas e Robótica & Instituto Superior Técnico
Portugal
{plinio,pedro,jasv}@isr.ist.utl.pt

Abstract. We address the empirical feature selection for tracker-less recognition of human actions. We rely on the appearance plus motion model over several video frames to model the human movements. We use the L_2 Boost algorithm, a versatile boosting algorithm which simplifies the gradient search. We study the following options in the feature computation and learning: (i) full model vs. component-wise model, (ii) sampling strategy of the histogram cells and (iii) number of previous frames to include, amongst others. We select the features' parameters that provide the best compromise between performance and computational efficiency and apply the features in a challenging problem, the tracker-less and detection-less human activity recognition.

1 Introduction

Works on human activity recognition rely on detection and tracking algorithm in order to discriminate the human patterns present in videos [9]. On one hand, the detection algorithms are image-based approaches that segment the region of interest for further processing [6]. On the other hand, tracking algorithms use the detector output and data association techniques to segment video regions where the activity patterns are learnt and matched (e.g. [1]).

The state-of-the-art approaches for people detection and tracking have attained very good performances in challenging data sets (see [16]). However, their application on more realistic scenarios does not provide good results yet due to the following challenges: real-time video stream input, outdoor illumination variations, large amounts of clutter, motion blur, moving cameras, amongst others. Since most of the human activity recognition approaches assume flawless detectors and trackers, their application on more challenging scenarios is even more difficult. Considering these constraints for the application of human

* This work was supported by FCT (ISR/IST plurianual funding through the PIDDAC Program) and partially funded by EU Project First-MM (FP7-ICT-248258), EU Project HANDLE (FP7-ICT-231640) and by the project CMU-PT/SIA/0023/2009 under the Carnegie Mellon-Portugal Program.

activity recognition on real scenarios, we address the following questions in this paper:

1. Is it possible to remove the tracking algorithm and find features for activity recognition with good performance, assuming a flawless person detector?
2. If (1) is possible, would the found features work properly in a scenario without detector? In other words, would be feasible to detect people and recognize their activities?

In order to address the questions above, we rely on the state-of-the-art model for human activity recognition: the combination of appearance and motion patterns of each activity [9]. The appearance is encoded by the histogram of image gradients and the motion is encoded by the histogram of the optic flow (dense). In order to learn how to discriminate the patterns we use the popular boosting algorithms, which are efficient, versatile and have shown similar recognition results to more elaborate techniques. Our choice is the L_2 Boost algorithm [3], which has two main differences with common boosting methods (e.g. AdaBoost): i) the data points do not have weights to adapt because they are basically included in the gradient computation and ii) the weak learners do not have weighting coefficients because L_2 boost uses a fixed step size equal to 1.

We choose the Weizmann dataset for the experiments, originally recorder by [2], because it addresses an interesting multiclass problem that has been virtually solved using the detector plus tracker assumption [7][11]. Thus, the common training and testing steps of the previous works use the the location and size of the people over time, provided by the groundtruth.

In order to address question (1) we use the location and size for each frame separately, so the temporal data association is not considered. We build a spatio-temporal cuboid for each detection independently, so the detected region of interest is projected onto the previous frames. This means that the person may not fully visible on the previous regions of interest. Then, the feature selection procedure searches for the parameters of feature computation that provide very good recognition results and low computational requirements.

In order to answer question (2), we use the features obtained in the previous step and add the “background class” (i.e Nobody performing any activity) to the multi-class problem. Thus, we are able to apply the sliding window method in order to detect people and recognize their activities. In the case of video sequences, the sliding window turns into the sliding cuboid for person and activity detection. The results show that the tracker-less activity recognition is plausible, while the tracker-less and detector-less activity recognition is a very difficult problem.

2 Human Activity Model

The state-of-the-art action recognition approaches use a combination of appearance and motion-based features in order to extract the activities’ patterns from videos [11]. We follow this approach, using the image gradient and optical flow (dense) as the raw features to extract the action patterns. Figure 1-A and 1-B

show an example of the video volume (cuboid) for feature computation. Note that the person's bounding box at frame I^t maintain the same location over the previous $\tau - 1$ frames, so we do not consider the data association provided by a tracking algorithm. Thus, we use only the person's location at the current frame, making the problem even more complicated, but allowing for an easier development toward the use of moving cameras (for instance mounted on moving robots). The most discriminative and efficient features based on gra-

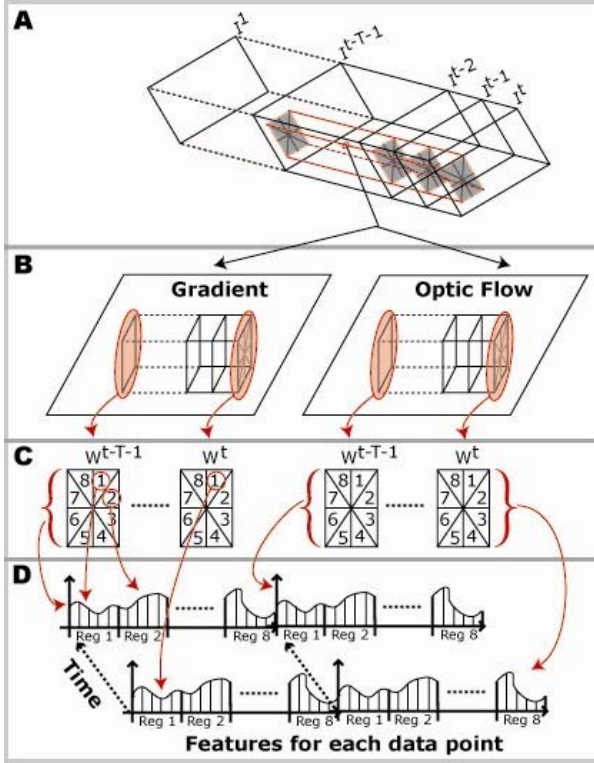


Fig. 1. Feature computation (extracted from [10]): A) example of a volume of video used to compute the features for the person detected in image I^t , B) the two types of raw features used, gradient and flow vectors, computed inside the volume correspondent to the person detected, C) polar sampling used to divide each window into subregions and D) weighted histograms computed for each region, producing a 2D matrix coding the evolution of each bin over a set of T frames

dients compute weighed histogram of the raw features, such as the histogram of gradients (HOG) [4] and histogram of optic flow [5]. Given a gradient image or optic flow image, the weighed histogram divides the image in subregions (according to a sampling strategy, e.g. Cartesian, polar) and computes the histogram of the gradient (or flow) orientation weighed by its magnitude. Figure

II-C shows the polar sampling strategy. In the case of polar sampling, the histogram features are parametrized by the number of subregions (cells) nR and the number of bins nB for each subregion. The correspondent parameters of Cartesian sampling, are the number of intervals the x direction nI_x , the number of intervals in the y direction nI_y and the number of bins nB , which defines $nI_x \times nI_y$ subregions (cells). We denote the gradient histogram as the row vector $g^t \in \mathbb{R}^{nB \cdot nR}$ and $g^t \in \mathbb{R}^{nI_y \cdot nI_x \cdot nB}$ for the polar and Cartesian histograms respectively. Similarly, the flow histograms are denoted as the row vector $o^t \in \mathbb{R}^{nB \cdot nR}$ and $o^t \in \mathbb{R}^{nI_y \cdot nI_x \cdot nB}$, computed at frame t .

At frame I^t and its correspondent rectangular region of interest $R(x_c, y_c, w, h)$ ¹, the appearance and motion feature vector for each person detected is

$$h^{t,R} = [g^t o^t] \in \mathbb{R}^{2 \cdot nB \cdot nR} \quad \text{polar sampling} \quad (1)$$

$$h^{t,R} = [g^t o^t] \in \mathbb{R}^{2nI_y \cdot nI_x \cdot nB} \quad \text{cartesian sampling} \quad (2)$$

We consider two ways of modeling the human activity patterns in the spatio-temporal cuboid: (i) the component-wise approach and the (ii) full representation. The component-wise stacks the vector component h_j^t in the previous $t + \tau - 1$ frames, so the row feature vector is as follows:

$$X_i^j = [h_j^t \dots h_j^{t+\tau-1}]. \quad (3)$$

The full representation stacks all the h^t vectors in the previous $\tau - 1$ frames,

$$X_i = [h^t \dots h^{t+\tau-1}], \quad (4)$$

where i is the data sample index.

3 L₂Boost with Temporal Models

The binary L₂boost algorithm estimates the function $F: \mathbb{R}^d \rightarrow \mathbb{R}$ by minimizing the expected cost $\mathbb{E}[C(y, F(X))]$ based on the data $(y_i, X_i), i = 1, \dots, n$. The cost function is $C(y, f) = (y - f)^2/2$ with $y \in \{-1, 1\}$ and its respective population minimizer is $F(X) = \mathbb{E}[y|X = x]$. The overall optimization is achieved by means of a sequential stagewise approximation along M rounds, optimizing a so called weak learner in each round, m [3]. The weak learner is the linear combination of the components of the feature vector X_i , so the weak learner of the component-wise model of Eq (3) is $f_m(X_i^j) = X_i^j \beta^m$ and for the full model of Eq. (4) is $f_m(X_i) = X_i \beta^m$.

In order to use matrix notation, we stack all the y_i values into the vector $Y \in \mathbb{R}^N$ and all the X_i data points into the matrix X . In the case of the component-wise model, at each round m , we optimize a temporal model β for each possible feature $j = 1, \dots, D$, choosing the one that achieve less error:

$$\hat{\beta} = \arg \min_{\beta, j} (Y - X^j \beta)^T (Y - X^j \beta). \quad (5)$$

¹ Centroid, width and height.

The solution is $\hat{\beta}^m = (X^{j^m T} X^{j^m})^{-1} X^{j^m T} Y$, where j^m is the component that achieves less error. In the case of the full model of Eq. (4), the feature index j is removed from Eq. (5), so $\hat{\beta} = \arg \min_{\beta} (Y - X\beta)^T (Y - X\beta)$, whose solution is $\hat{\beta}^m = (X^T X)^{-1} X^T Y$. The component-wise L_2 boosting algorithm with linear temporal models of Eq. (3) is as follows:

1. **Initialization.** Chose M and set $m=0$. Given data (Y, X) , fit the first weak learner, $\hat{F}_0 = X^{j^0} \hat{\beta}^0$. β^0 and j^0 are computed from Eq. (5).
2. **Projection of gradient to learner.** Compute the negative gradient (in this case are the residuals) $u_i^{m+1} = y_i - \hat{F}_m(X_i)$ ($i = 1, \dots, n$). For simplicity, stack all u_i values into the vector $U \in \mathbb{R}^N$. Use the residuals U^{m+1} to fit the learner $\hat{f}_{m+1} = X^{j^{m+1}} \hat{\beta}^{m+1}$ changing Y for U in Eq. (5). Update $\hat{F}_{m+1} = \hat{F}_m + \hat{f}_{m+1}$. Compute $\hat{F}_{m+1} = \text{sign}(\tilde{F}_{m+1}) \min(1, |\tilde{F}_{m+1}|)$.
3. **Iteration.** If $m + 1 < M$ increase m by 1 and goto step2. If $m + 1 = M$ return $\Theta_j = \{j^0, \dots, j^m, \dots\}$, and one set of models, $\Theta_{\beta} = \{\beta^0, \dots, \beta^m, \dots\}$

The classification of a new point X_i is given by the sign of the strong classifier result, $\text{sgn} \hat{F}_M(X_i)$. Notice that the last computation of step 2 constraints the strong classifier to be in $[-1, 1]$, so we apply the L_2 Boost with constraints [3], which works better in the classification setup. The algorithm just presented is very similar to the full model one, but removing the feature index j . The strong classifier $F(x)$ relates the class-conditional probabilities,

$$F(x) = 2p(y = 1|x) - 1, |F(x)| = |p(y = 1|x) - p(y = -1|x)|, \quad (6)$$

and its module $|F(X_i)|$ is the classification margin, that is the probability of labeling the new data point given the models estimated. In order to extend the L_2 Boost with linear-temporal models to multi-class problems we use the one vs. all approach, which solves C binary problems to discriminate between C classes where $Y \in \{1, \dots, C\}$. The multi-class version of L_2 starts by computing $\hat{F}_M^{(c)}$ on the basis of the binary response variables

$$Y_i^{(c)} = \begin{cases} 1 & \text{if } Y_i = c \\ -1 & \text{if } Y_i \neq c \end{cases} \quad i = 1, \dots, n \quad (7)$$

and then builds the classifier as $\hat{C}^m(x) = \arg \max_{c \in \{1, \dots, C\}} \hat{F}_M^{(c)}(x)$.

4 Feature Selection for Tracker-Less Recognition

We address this problem by comparing the recognition rate between different types of features in the Weizmann dataset [2], which contains 9 subjects

performing 9 actions: {1 - bending down, 2 - jumping jack, 3 - jumping, 4 - jumping in place, 5 - running, 6 - galloping sideways, 7 - walking, 8 - waving one hand, 9 - waving both hands}. We follow the evaluation protocol proposed by [2] that performs a leave-one-out test with the 9 subjects, so each subject belongs to one of the testing sets. Then, the confusion matrix is averaged over all the leave-one-out test sets and the trace of the averaged matrix is used as the measure of recognition performance.

We consider the following options to select the feature computation method: (i) component-wise vs. full model, (ii) cartesian and polar cell sampling, (iii) number of frames τ of the linear temporal model, (iv) optic flow algorithm, (v) two options for the region of interest in the image (detected bounding box) and (vi) cell overlapping. We observe in Table 1 that the component-wise L_2 Boost performs better than the full model one, so in the rest of the experiments we just consider the component-wise approach. In addition, we select the polar and cartesian sampling that attain the top recognition result, $nR = 16$, $nB = 16$ for polar and $nI_x = 4$, $nI_y = 8$ for cartesian. The next step is to compare the effect of the optic flow algorithm in the classification results. Table 2 shows that Ogale's et. al. [8] algorithm has a better performance than Werlberger's one. In this case our choice is the Werlberger's algorithm because of the GP/GPU implementation that allows to compute the optic flow (dense) in near real-time for normal cameras. The reason behind this choice is the quicker evaluation of our approach on other datasets (e.g. [10]), and the near real-time plausibility of

Table 1. Component-wise vs. full model results, using two sets of parameters for each sampling approach. ($\tau = 10$, no overlapping between cells and using groundtruth detections), Ogale's optic flow [8].

Feature type	Average confusion matrix's trace (%)		dims
	component-wise	all features	
polar $nR = 8$, $nB = 16$	91,29	89,76	256
polar $nR = 16$, $nB = 16$	95,42	93,2	512
cartesian $nI_x = 4$, $nI_y = 8$, $nB = 16$	95,42	93,2	512
cartesian $nI_x = 3$, $nI_y = 6$, $nB = 16$	95,46	92,79	576

Table 2. Effect of two optic flow approaches on the recognition rate ($\tau = 10$, no overlapping between cells and using groundtruth detections)

Feature type	Average confusion matrix's trace (%)	
	Ogale et. al. [8]	Werlberger et. al. [12]
polar $nR = 16$, $nB = 16$	95,42	94,11
cartesian $nI_x = 4$, $nI_y = 8$, $nB = 16$	96,01	94,9

[12], which facilitates future deployment of the system. The temporal support used in the previous test ($\tau = 10$) was motivated by Schindler et. al. [11]. Table 3 re-validates their choice $\tau = 10$. In the following we compare the groundtruth boxes against a manually set bounding box for all the detections. We define a bounding box with constant width/height ratio in order to select the spatio-temporal cuboids. The rationale of this fixed ratio bounding box is two-folded: (i) facilitate the application of the sliding window method and (ii) allow the search over multiple scales. Table 4 shows that the selected w/h is practically equal to the groundtruth boxes, because the persons of the Weizmann dataset have similar sizes. Finally, we apply the idea of overlapping between cells [4]. In the case of polar sampling, we add more cells to the previous ones in such a way that each new cell overlaps with two of the original neighboring cells in equal proportion. In the case of the cartesian sampling, each new cell overlaps with four of the original neighboring cells in equal proportions. Table 5 shows that cell overlapping and cartesian sampling brings better results, but at the expense of a larger computational load. Since we are interested in features having a lower computational load and good performance, we choose the polar sampling with no overlap. Summarizing, the feature selection options are: (i) component-wise L_2 Boost, (ii) Welberger’s optic flow [12], (iii) $\tau = 10$, (iv) fixed w/h ratio bounding boxes and (v) no overlap polar sampling cells.

Table 3. Temporal support comparison. (no overlapping between cells and using groundtruth detections).

τ	1	3	5	7	10	13	15
polar $nR = 16, nB = 16$	86,2	90,36	92,64	93,27	94,11	93,55	93,47
cartesian $nI_x = 4, nI_y = 8, nB = 16$	87,88	91,7	92,35	94,1	94,9	94,36	94,11

Table 4. Region of interest comparison. Groundtruth boxes vs. manually selected ones. (no overlapping between cells).

Feature type	Average confusion matrix’s trace (%)	
	groundtruth ROI [2]	Fixed size ROI $w = 60, w/h = 0.779$
polar $nR = 16, nB = 16$	94,11	94,84
cartesian $nI_x = 4, nI_y = 8, nB = 16$	94,9	95,56

Table 5. Comparison between cells with and without overlap

Feature type	Average confusion matrix’s trace % (dimensions)	
	half interval overlap	no overlap
polar $nR = 16, nB = 16$	95,15 (1024)	94,84 (512)
cartesian $nI_x = 4, nI_y = 8, nB = 16$	95,68 (1696)	95,56 (1024)

4.1 Tracker-Less and Detection-Less Scenario

The features found above attain very good recognition rates in a tracker-less scenario. In this section we want to evaluate their performance on a tracker-less and detector-less scenario. Thus, we need to add the background activity class (i.e. spatio-temporal cuboids where no person is doing any action) to the activity classes in order to both detect people and recognize their activities. We obtain the background samples by the random selection of video segments in the Weizmann dataset. Then, we compute the features selected in the previous section and re-train the L_2 Boost algorithm with the 9 activities plus the “background” activity.

The testing phase comprises the application of the volume-based version of the sliding window algorithm. This image-based algorithm is applied on pedestrian detection, by moving the region of interest (window) along the image grid. For each grid point, the image features are computed inside the window, followed by the binary classification (person or background). We perform the volumetric version of the algorithm, by moving the region of interest (cuboid) along the video (3D) grid. Then, we classify each cuboid as a particular human activity or the “background” activity. We sample the video grid every 5 pixels in each image direction and every 2 frames in the temporal direction. The trace of the confusion matrix for the 10 classes is 94, 74%. This looks like a good result because of the larger number of background samples compared to the human activity samples. After removing the background samples the trace of the confusion matrix is 30, 04%. This result is explained by the absence of the perfectly aligned detection results provided by the groundtruth. These misalignments of the cuboids in the video were not learnt during training, so the L_2 Boost is not able to discriminate between the human activities.

5 Conclusions

We address the feature selection for human activity recognition in a tracker-less scenario. We construct features that encode appearance and motion by means of the Histogram Of Gradients (HOG) [4] and the Histogram Of Flow (HOF) [5] over several video frames. Our choice of learning approach, the L_2 Boost, it finds the linear models for binary problems and we apply the one vs. all approach for the final classification.

In this feature-classifier context, we select experimentally the parameters that: (i) attain very good results and (ii) have low computational requirements. In addition, we evaluate the selected features in a tracker-less and detector-less scenario, a very challenging problem due to the large appearance variation in the background and the reduced amount of motion information contained in it. Future work must study the combination of features from both worlds: human activity recognition and pedestrian detection in order to have features that do not assume flawless person trackers and detectors.

References

1. Andriluka, M., Roth, S., Schiele, B.: People-tracking-by-detection and people-detection-by-tracking. In: IEEE CVPR 2008, pp. 1–8 (2008)
2. Blank, M., Gorelick, L., Shechtman, E., Irani, M., Basri, R.: Actions as space-time shapes. In: IEEE ICCV 2005, vol. 2, pp. 1395–1402 (2005)
3. Buhlmann, P., Yu, B.: Boosting with the l2 loss: Regression and classification. *Journal of the American Statistical Association* 98, 324–339 (2003)
4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Proceedings of the CVPR 2005, Washington, DC, USA, pp. 886–893 (2005)
5. Dalal, N., Triggs, B., Schmid, C.: Human detection using oriented histograms of flow and appearance. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3952, pp. 428–441. Springer, Heidelberg (2006)
6. Gerónimo, D., López, A., Sappa, A., Graf, T.: Survey of pedestrian detection for advanced driver assistance systems. *IEEE PAMI* 32(7), 1239–1258 (2010)
7. Jhuang, H., Serre, T., Wolf, L., Poggio, T.: A Biologically Inspired System for Action Recognition. In: Proceedings ICCV, pp. 1–8 (October 2007)
8. Ogale, A.S., Aloimonos, Y.: A roadmap to the integration of early visual modules. *International Journal of Computer Vision* 72(1), 9–25 (2007)
9. Poppe, R.: A survey on vision-based human action recognition. *Image and Vision Computing* 28(6), 976–990 (2010)
10. Ribeiro, P.C., Moreno, P., Santos-Victor, J.: Unsupervised and online update of boosted temporal models: the UAL₂boost. In: Proc. of ICMLA (December 2010)
11. Schindler, K., van Gool, L.: Action snippets: How many frames does human action recognition require? In: IEEE CVPR 2008, June 2008, pp. 1–8 (2008)
12. Werlberger, M., Trobin, W., Pock, T., Wedel, A., Cremers, D., Bischof, H.: Anisotropic Huber-L1 optical flow. In: Proc. of BMVC (September 2009)

Classification of Atomic Density Distributions Using Scale Invariant Blob Localization

Kai Cordes¹, Oliver Topic², Manuel Scherer²
Carsten Klempt², Bodo Rosenhahn¹, and Jörn Ostermann¹

¹ Institut für Informationsverarbeitung (TNT), Leibniz Universität Hannover

<http://www.tnt.uni-hannover.de>

² Institut für Quantenoptik (IQO), Leibniz Universität Hannover

<http://www.iqo.uni-hannover.de>

Abstract. We present a method to classify atomic density distributions using CCD images obtained in a quantum optics experiment. The classification is based on the scale invariant detection and precise localization of the central blob in the input image structure. The key idea is the usage of an a priori known shape of the feature in the image scale space. This approach results in higher localization accuracy and more robustness against noise compared to the most accurate state of the art blob region detectors.

The classification is done with a success rate of 90% for the experimentally captured images. The results presented here are restricted to special image structures occurring in the atom optics experiment, but the presented methodology can lead to improved results for a wide class of pattern recognition and blob localization problems.

1 Introduction

1.1 Atomic Density Distributions

Satyendranath Bose and Albert Einstein predicted in 1924 that a gas of atoms with integer spin forms a so-called Bose-Einstein condensate (BEC) when it is cooled to ultra cold temperature [1]. Below a certain temperature threshold, a large fraction of atoms confined in an external trap occupy the physical ground state. In 1995, two experimental groups achieved Bose-Einstein condensation of trapped dilute atomic gases [2,3] after cooling it below a temperature of $1 \mu\text{K}$. At these temperatures, the velocity distribution is very narrow and due to Heisenberg's uncertainty relation [4] the spatial distributions of those atoms is broad. Typically, it extends to several tens of micrometers, making it possible to image the ensemble with a CCD camera as shown in Fig. 1.

In the past years, BEC's consisting of atoms with non-zero spin attracted a lot of notice. These atoms behave like little magnets that may be oriented perpendicular to an external magnetic field. Atomic collisions can now generate pairs of atoms, with one spin pointing upwards and the other one downwards. This process was identified to be a parametric amplifier for classical seed atoms

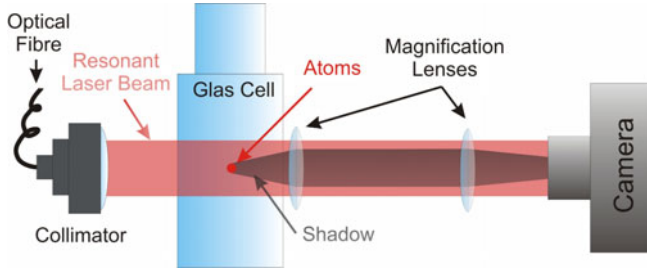


Fig. 1. Simplified sketch of the imaging system. The atomic cloud is illuminated by collimated resonant laser light from an optical fibre. The shadow from the atoms is imaged by a magnifying lens system onto a CCD-Camera. Subtraction from an image without atoms leads to the atomic density distribution.

or vacuum fluctuations [5,6]. Depending on the magnetic field, those atoms can be generated in different states with different characteristic probability distributions [7]. In the cylindrical trapping geometry used in the recent experiments, the discrete physical states that may be populated have density distributions $n(r)$ which can be approximated with the following expression

$$n_{nl}(\mathbf{r}) \propto J_l^2(\beta_{nl} \frac{|\mathbf{r}|}{r_{\text{tf}}}) \quad (|\mathbf{r}| < r_{\text{tf}}), \quad (1)$$

where J_l are the Bessel functions of the first kind and β_{nl} is the n th zero of J_l . The size is scaled by the radius r_{tf} . Each distribution is identified by two quantum numbers n and l for the radial excitation and the rotation of the cloud.

After preparing the clouds, the trap is switched off to allow for ballistic expansion, where the distribution is stretched but not perturbed [8]. During the expansion, the three spin components are separated by an applied magnetic field gradient and then irradiated with a resonant laser beam. The atomic clouds absorb light and the resulting shadow is imaged onto a CCD camera. From the CCD data, the density distribution can be determined. The imaging technique has three deficiencies: Interferences in the detection beam produce regular stripes on the density distribution. The imaging setup can distort the image slightly. The finite number of detected photons leads to shot noise on the images.

In general, we detect pictures with clouds in arbitrary combinations of quantum numbers. Fig. 2 shows the three examples under investigation (I_0 , I_1 , and I_2) with the quantum numbers $(n, l) = (1, 0)$, $(2, 0)$, and $(2, 1)$. For the interpretation of the experimental results, an unambiguous classification of the quantum numbers is of key interest. This classification should be independent of the total position and the size of the clouds, since these parameters are quickly changed by a variation of the experimental parameters and technical uncertainties. Additionally, the classification should be robust with respect to the experimental noise on the figures. In the following, we describe how the density distribution can be classified automatically and the quantum numbers are inferred.

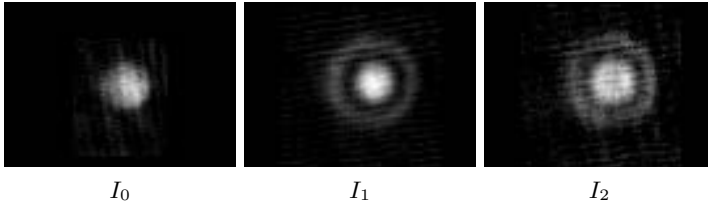


Fig. 2. Real atomic distribution shapes to be detected and classified. From left to right: type I_0 , I_1 , and I_2 . *Best viewed on a LCD.*

1.2 Feature and Blob Detectors

The objective of this work is to classify the three different atomic distribution shape types I_0 , I_1 , and I_2 as shown in Fig. 2. For each distribution shape, the underlying function is known from equation (II). The shapes I_0 and I_1 are identical (proportional to $J_0^2(\cdot)$), but differ by the radial excitation. The shape of I_2 is proportional to $J_1^2(\cdot)$. As the size of the blobs can vary, it is necessary to perform a scale invariant classification. Due to the currently small available data set, a training scheme for the classification is not applied and the proposed approach concentrates on feature estimation. The **contributions** are:

- a new detector robust to noise for the localization of one unique feature of known shape with high accuracy,
- the comparison of the detector to the most accurate state of the art blob detectors using synthetic images, and
- the classification of atomic distribution shapes using the extracted shape parameters in a unified feature detection framework.

The initial and most important task is the accurate localization of the central blob in the images. Then, the classification can be done in two steps. First, the type of extremum in the input image is determined to separate the types I_0 , I_1 from I_2 . Second, the ring surrounding the blob is localized and used for the separation of type I_0 (no ring visible) and I_1 (ring is visible).

Due to noise and the imaging setup, the target structure might be slightly slanted. Thus, the desired method for the detection task has to be an affine invariant noise resistant blob region detector. In literature, extensive work has been done on region detectors and their evaluation. An overview of most of these detectors can be found in [9], in which the most accurate affine invariant blob detectors are found to be the Hessian-Affine [10] and the MSER [11]. Their evaluations show excellent performance [9,12] regarding the *Repeatability* rate, which is the most often used criterion for elliptical region localization accuracy. In [11], maximally stable extremal regions (MSER) are constructed using a segmentation process. Then, an ellipse is fit to each of the detected regions. Based on affine normalization, the Hessian-Affine detector determines the elliptical shape with the second moment matrix of the intensity gradient. The features are detected as extrema in the scale space, which is introduced and described by Lindeberg et al. [13,14]. The scale space representation is built by cascading

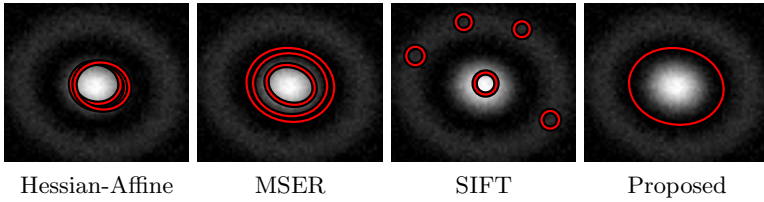


Fig. 3. Localization results of state of the art blob detectors. From left to right: Hessian-Affine, MSER, SIFT, and the proposed approach which aims to localize the first zeros of the input feature. *Best viewed on a LCD.*

Gaussian filters of differing standard deviation σ . The scale space is also used by the SIFT detector [15] as the basis for blob detection. In SIFT, the Difference of Gaussians (*DoG*) pyramid is evaluated as an approximation of the scale space of the input image. A scale space extremum is detected as a luminance value that is bigger or smaller than its 26 neighbors in the *DoG* pyramid. Although the SIFT detector is not affine invariant by design, it shows impressive performance for features with moderate affine distortion. In [16], the localization accuracy of SIFT is increased using a bivariate approximation of the image gradient signal. This approach is adapted for the localization of the feature shapes occurring in the atom optics experiment.

Results of the state of the art blob detectors for an example are shown in Fig. 3. While the Hessian-Affine, MSER, and SIFT detectors lead to ambiguous results, the desired method provides an unique and accurate detection of the central blob, which is defined by the bounding zeros.

Our work demonstrates, that the elliptical shape of these features can be determined with high accuracy by incorporating the knowledge of the underlying functions. It is shown that all three input feature types (Fig. 2) approximately depict the same shape in the dominant scale in the scale space, which leads to a unified detection and localization procedure. Incorporating shape knowledge of the input data, the ring area surrounding the center blob can be determined and used for the classification. In the following Section 2, the approach of localization and classification of the distribution shapes is presented. Section 3 shows experimental results using synthetically constructed and real image data. In Section 4, the paper is concluded.

2 Localization and Classification of Atomic Density Distribution Shapes

A blob feature as shown in Fig. 3 is defined by image coordinates (x_0, y_0) , and the covariance matrix $\Sigma = \begin{pmatrix} a^2 & b \\ b & c^2 \end{pmatrix}$, which determines the elliptical shape. In order to estimate these parameters of the input feature, it has to be detected and localized in the scale space. Here, the Difference of Gaussians (*DoG*) representation is

used, which is a good approximation as proved by the SIFT approach [15]. An experimental analysis using the SIFT scale selection technique (Section 2.1) leads to the proposed function model which approximates the image signal in the selected scale of the *DoG*. The function model used for the localization is explained in Section 2.2. A robust technique for the detection and localization is derived in Section 2.3. On basis of the extracted localization parameters, the classification of the atomic distributions is done as explained in Section 2.4.

2.1 Feature Shape in the Scale Space

The feature selection scheme of the SIFT detector provides the best representation of a feature in the scale space. The scale is determined by the octave *o* and the interval *i* [15]. In Fig. 4, the selected scales for the synthetic input features \tilde{I}_0 , \tilde{I}_1 , and \tilde{I}_2 (top row) are shown in the bottom row. The returned shapes of the input features \tilde{I}_0 , \tilde{I}_1 , \tilde{I}_2 are approximately sinc functions. This observation leads to the assumption for the approximation of a feature in the scale space as shown in the following Section 2.2. Note, that the central blob in the input images lead to minimas in the scale space for all feature types $\tilde{I}_0, \tilde{I}_1, \tilde{I}_2$.

2.2 Localization Using the SINC Function Model

Following the observation that the returned shapes in the Difference of Gaussians pyramid are approximately sinc functions (Section 2.1), the input features

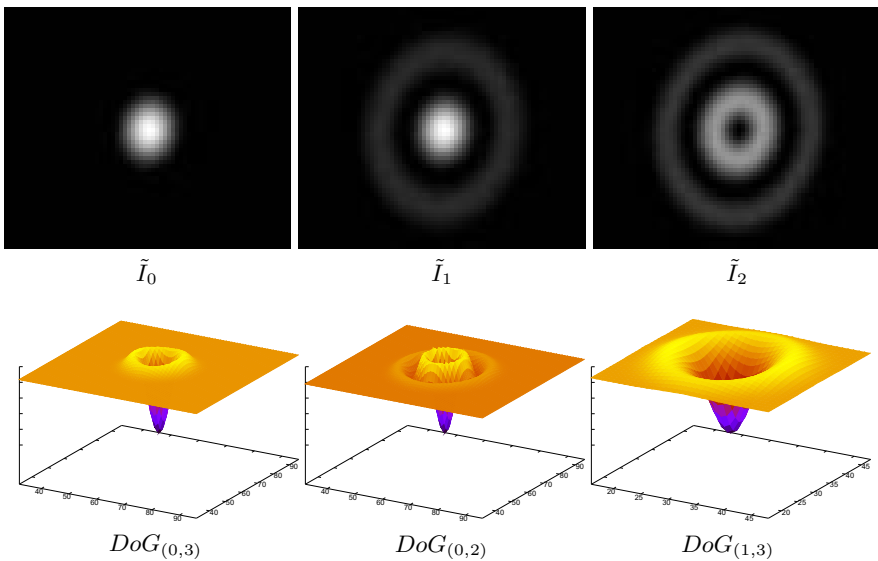


Fig. 4. Resulting image signal $DoG_{(o,i)}$ of interval *i* in octave *o* using the scale selection of SIFT. The synthetic test images are shown on top. Each of the input features $\tilde{I}_0, \tilde{I}_1, \tilde{I}_2$ depict a sinc $(r) = \frac{\sin r}{r}$ shape in the selected scale (bottom).

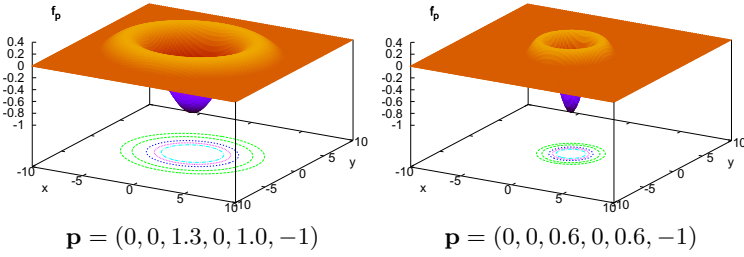


Fig. 5. Proposed regression function $f_{\mathbf{p}}(\mathbf{x})$ for the approximation of the scale space shapes of the input blobs as shown in Fig. 4. Two examples with different covariance matrix Σ are shown.

are localized using this function model. To allow elliptical feature shapes, the covariance matrix $\Sigma = \begin{pmatrix} a^2 & b \\ b & c^2 \end{pmatrix}$ is incorporated. For the following, the abbreviation $R_{\mathbf{x}_0, \Sigma}(\mathbf{x}) := (\mathbf{x} - \mathbf{x}_0)^\top \Sigma^{-1}(\mathbf{x} - \mathbf{x}_0)$ is used. $R_{\mathbf{x}_0, \Sigma}(\mathbf{x})$ is used to describe the elliptical shape with the center coordinate $\mathbf{x}_0 = (x_0, y_0)$. Together with a peak value v , the parameter vector $\mathbf{p} = (x_0, y_0, a, b, c, v)$ determines a member of the following proposed function model $f_{\mathbf{p}}$ for the detection and localization approach:

$$f_{\mathbf{p}}(\mathbf{x}) = \begin{cases} v \cdot \frac{\sin \sqrt{R_{\mathbf{x}_0, \Sigma}(\mathbf{x})}}{\sqrt{R_{\mathbf{x}_0, \Sigma}(\mathbf{x})}}, & \text{for } R_{\mathbf{x}_0, \Sigma}(\mathbf{x}) \leq t_0 \\ 0 & , \text{ otherwise} \end{cases} \quad (2)$$

with $t_0 = 2\pi$. Note, that the peak value v has to be negative $v < 0$ to detect the desired extremum (see Fig. 4). Scale space maxima are not considered for the localization. Two examples for the function model $f_{\mathbf{p}}$ are shown in Fig. 5. They are determined by the parameter vector $\mathbf{p} = (x_0, y_0, a, b, c, v)$ with six components. The parameter vector \mathbf{p} of an input feature is identified by means of a regression analysis. Each fullpel position in each octave o and each interval i is assumed as a possible initialization for the Levenberg-Marquardt optimization algorithm. The covariance matrix is initialized with the unit matrix $\Sigma = \mathbf{E}$, which is equivalent to circular shape. As each scale is normalized in the DoG pyramid, the initial value for v is -1 . The Levenberg-Marquardt (LM) algorithm minimizes the distance $e_{\mathbf{p}}$ between the model function $f_{\mathbf{p}}(\mathbf{x})$ and the image signal $DoG_{(o,i)}(\mathbf{x})$ in the current scale (o, i) evaluating a squared neighborhood \mathcal{N} :

$$e_{\mathbf{p}} = \sum_{\mathbf{x} \in \mathcal{N}} (f_{\mathbf{p}}(\mathbf{x}) - DoG_{(o,i)}(\mathbf{x}))^2 \quad (3)$$

The LM algorithm stops returning the optimal parameter vector \mathbf{p}_{opt} and a residuum value $e_{\mathbf{p}_{opt}}$ which provides a quality measure of the obtained regression function $f_{\mathbf{p}_{opt}}(\mathbf{x})$ for the initial starting position.

In contrast to the SIFT detector, our approach using a regression analysis is capable of evaluating arbitrary neighborhood sizes. As can be seen in Fig. 4, a

neighborhood \mathcal{N} of at least 9×9 pixels is needed to capture the characteristics of the blob shape in the image pyramid. A large neighborhood also leads to a localization which is less sensitive to noise. The SIFT detector uses a 3×3 neighborhood and the neighboring scales to determine the subpel and subscale localization. To compensate for the computational expense of a larger neighborhood, our approach omits to estimate a subscale parameter.

2.3 Feature and Scale Selection

The regression analysis described in Section 2.2 returns a residuum $e_{\mathbf{p}_{opt}}$ for the optimal parameter vector \mathbf{p}_{opt} , which is a quality measure for the resulting regression function $f_{\mathbf{p}_{opt}}(\mathbf{x})$. Hence, the best blob location is found by minimizing the residuum for all possible positions. For the detection of the central blob, it is crucial to favor a scale space minimum in smaller scales. Therefore, the function to be minimized is weighted by the scale $w_{scl} = 2^{o+\frac{k}{2}}$, where k is the number of scales per octave [15] (usually $k = 3$):

$$w_{scl} \cdot e_{\mathbf{p}_{opt}} \rightarrow MIN \quad (4)$$

To ensure optimal solutions, a brute force search is performed within a search range. The brute force search and the large neighborhood lead to a significant increase in computational complexity. This is not critical for the presented classification application.

2.4 Classification of the Feature Shapes

The classification workflow is shown in Fig. 6. Two evaluations are done after localizing the best feature blob. First, the feature type I_2 is distinguished from the others by determining the **Curvature** of the input image at the localized position. This is done by evaluating the first scale of the *DoG* pyramid at the ground plane position $\mathbf{x}_G = \mathbf{x}_0 \cdot 2^o$. To reduce the influence of noise, the median \mathcal{N}_{med} in a 3×3 neighborhood \mathcal{N} is used to classify between $I_0 \cup I_1$ (concave curvature) and I_2 (convex curvature):

$$\mathcal{N}_{med}(DoG_{(0,0)}(\mathbf{x}_G)) \leq 0 \Rightarrow I_0 \cup I_1 \quad (5)$$

$$\mathcal{N}_{med}(DoG_{(0,0)}(\mathbf{x}_G)) > 0 \Rightarrow I_2 \quad (6)$$

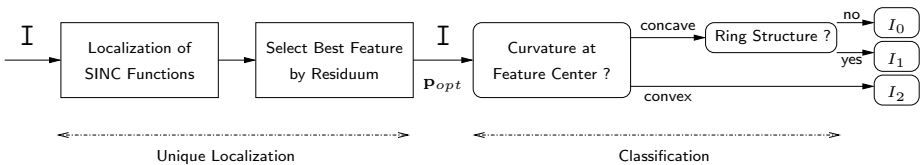


Fig. 6. Workflow diagram of localization and classification of the input image I

The classification between I_0 and I_1 is done by evaluating if there is a **Ring Structure** around the center blob or not. The Ring \mathcal{S} is localized as the region between the first zeros z_{01} and the second zeros z_{02} of the Bessel function $J_0(\cdot)$:

$$\mathcal{S} = \{ \mathbf{x} : z_{01} \leq \sqrt{\frac{R_{\mathbf{x}_G, q \cdot \Sigma_G(\mathbf{x})}}{D(q \cdot \Sigma_G)}} \leq z_{02} \} \tag{7}$$

where $\Sigma_G = \Sigma \cdot 2^\circ$ is the covariance matrix of the feature in the image ground plane and $D(\cdot)$ denotes the determinant. The zeros of $J_0(\cdot)$ are known as $z_{01} \approx 2.40$ and $z_{02} \approx 5.52$. The ellipse scaling factor $q \approx 1.59$ maps the minima of the Bessel function $J_0(\cdot)$ to its first zeros and is calculated as the quotient of the first minimum and the first zero z_{01} of $J_0(\cdot)$. The region \mathcal{S} is localized after the central blob is accurately determined by \mathbf{p}_{opt} .

An example of the elliptical ring \mathcal{S} using relation (7) is shown in Fig. 9. Using this area, the two feature types I_0 and I_1 can be distinguished by analyzing the gray values inside \mathcal{S} . Therefore, the energy E_S of the image signal $I(\mathbf{x})$ inside the Ring \mathcal{S} is calculated and a threshold classifier with threshold thr is applied. The area A_S of the Ring \mathcal{S} is used for the normalization of E_S to obtain scale invariant energy values:

$$E_S = \frac{1}{A_S} \int_S |I(\mathbf{x})|^2 d\mathbf{x} \tag{8}$$

If $E_S < thr$, then the feature is of type I_0 , otherwise it is of type I_1 . The threshold thr can be chosen between 10 and 25 which is valid for all the experimental feature data in this paper as shown in Fig. 8.

3 Experimental Results

For the evaluation of our method, synthetic and real data is used. For the synthetic data, the ground truth localization of each feature is known. The detection accuracy of position $\mathbf{x}_G = (x_G, y_G)$ and shape Σ_G is shown using the *Surface Error* measure [10]. The *Surface Error* is a percentage value that is minimal if a detected ellipse area is exactly matching the ellipse determined by the ground truth values. The evaluation for the three different types of synthetic input features $\tilde{I}_0, \tilde{I}_1, \tilde{I}_2$ with added Gaussian noise is shown in Section 3.1. The spatial neighborhood \mathcal{N} evaluated for the feature localization is set to 13×13 pixels. For the real data, classification results of a set of images captured in the atom optics experiment are shown in Section 3.2. The processing time for an image (size 128×128) on common PC hardware is about 10 seconds, which is not critical for an automatic evaluation.

3.1 Results of Synthetic Data

Synthetic test images of types \tilde{I}_0, \tilde{I}_1 , and \tilde{I}_2 as shown in Fig. 4 (top row) are constructed using equation (1) and a cutoff at the first zeros for \tilde{I}_0 and second zeros for \tilde{I}_1, \tilde{I}_2 , respectively. For the evaluation, the following variations of the image signal are generated:

- scale s : $2 \leq s \leq 9$ (3 octaves) with step size 0.5
- subpel position x_0 : $-0.5 \leq x_0 < 0.5$ with step size 0.04
- noise variance σ_n : $0 \text{ dB} \leq \sigma_n \leq 80 \text{ dB}$ with step size 20 dB

Each of the variations has an impact on the localization accuracy. The subpel variation is to emphasize the signal approximation scheme used by the detectors while the scale variation emphasizes the scale invariance. The noise scenario demonstrates the robustness against image noise. For each synthetically constructed image, the feature is slightly slanted using a covariance Matrix Σ with $\frac{a}{c} = 1.2$.

For the classification task, it is crucial to select one unique feature for further evaluation, which is done by our method by design. The numbers of features selected by the presented approaches are shown in Fig. 7, top row. For the other detectors the numbers depend on the feature type and on the noise level. The localization accuracy evaluation is shown in Fig. 7, bottom row. If multiple features are detected by a method, the best of them is chosen for the evaluation. The cases in which no feature is detected are discarded from this evaluation (small scales for *Hessian-Affine*). To avoid the dependency on a global scale of the features, a normalization is applied to the covariance matrix results.

Our approach provides an accurate and reliable localization for each feature type compared to the best possible result of each of the other detectors. Due to the good subpel localization estimation, the SIFT detector provides comparably accurate results, but strongly increasing numbers of features with increasing noise. Interestingly, the localization accuracy of each detector does not increase significantly with increasing Gaussian noise. The detectors *Hessian-Affine* and

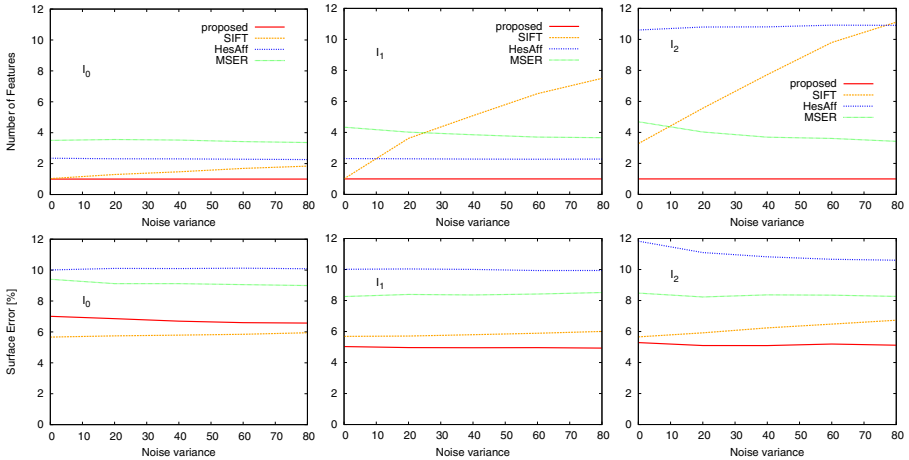


Fig. 7. Comparison of the number of detected features (top row) and the mean *Surface Error* (bottom row) for the three synthetic test features and the four region localization methods. In case of ambiguous detection results, the best is chosen for the *Surface Error*. From left to right: feature types I_0 , I_1 , and I_2

MSER result in highest Surface Errors. We can state that our results provide high accuracy which is robust to synthetic Gaussian image noise.

3.2 Results of Atom Optics Experiment Image Data

The captured image data include the real atom distributions resulting from the quantum optics experiment. To obtain equally distributed scales of features, additional input images are generated by resizing the original data set. To verify the thresholding approach explained in Section 2.4, the energy values E_S in equation (8) are shown in Fig. 8. Obviously, the types I_0 and I_1 are classified reliably and independently from the detected scale.

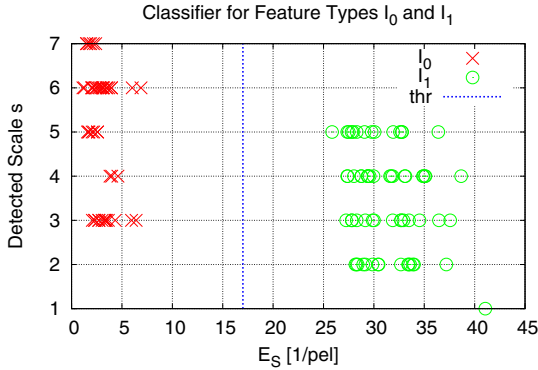


Fig. 8. Energy values E_S for the real experiment data for the feature types I_0 and I_1 of different sizes. It is shown that a simple threshold value thr is sufficient to separate the two classes. The classification is independent of the detected scale.

For the evaluation of the classification, 52 images of each type I_0 , I_1 , and I_2 are available. Examples are shown in Fig. 2. The classification rates for each input feature type and the two classification stages are shown in Table 1. The results for TP_{Curv} and TP_{Ring} demonstrate that misclassifications are only resulting from the curvature estimation in which only a small neighborhood is evaluated. Thus, this evaluation is more sensitive to the strong noise, especially for the input feature type I_2 . The *Ring Structure* detection and evaluation works perfect.

The overall correct classification rate is 90.4%. Classification failures are due to strong noise covering the center shape. In this case, blobs of type I_2 are

Table 1. Classification rate *True Positives* for the two stages *Curvature* TP_{Curv} and *Ring Structure* TP_{Ring} (see Section 2.4) and the resulting classification rate TP_{Σ}

	I_0	I_1	I_2	Σ
TP_{Curv}	96.2%	75.0%	85.6	
TP_{Ring}	100%	100%	–	100%
TP_{Σ}	96.2%	100%	75.0%	90.4%

very similar to the type I_1 (see Fig. 2). Understanding and modeling the noise structure, i.e. regular stripes from laser beam interferences, will improve the classification and is left for future works. Examples of the detected *Ring Structure* \mathcal{S} which is used to classify the features types I_0, I_1 are shown in Fig. 9.

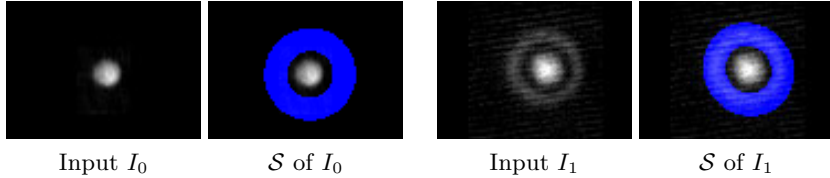


Fig. 9. Examples of the detected *Ring Structure* \mathcal{S} (in blue) around the central blob for two experimentally captured input images. *Best viewed on a LCD.*

4 Conclusion

The presented method consists of the detection, localization, and classification of atomic distribution shapes resulting from three types of modes from a quantum optics experiment. Therefore, a new feature detector is developed based on the SIFT approach. The a priori known shapes of the input features are incorporated using a regression analysis with a derived function model for the gradient signal. The determination of the function model parameters leads to a reliable and accurate localization of the elliptical shape of a feature blob. The shape parameters are used as input data for a simple two stage classifier.

The presented detector shows superior localization accuracy and noise robustness compared to the most accurate state of the art blob detectors. This is demonstrated using synthetic images. The classification success rate is 90% for the real data resulting from the atom optics experiment.

Our approach provides a useful application of scale invariant feature localization in the field of quantum optics. Future works will incorporate the noise structure for further classification improvements.

We acknowledge support from the Centre for Quantum Engineering and Space-Time Research QUEST.

References

1. Einstein, A.: Quantentheorie des einatomigen idealen gases. In: Sitzungsberichte der Preußischen Akademie der Wissenschaften Physikalisch-mathematische Klasse, pp. 261–267 (1924)
2. Anderson, M.H., Ensher, J.R., Matthews, M.R., Wieman, C.E., Cornell, E.A.: Observation of bose-einstein condensation in a dilute atomic vapor. *Science* 269, 198–201 (1995)
3. Davis, K.B., Mewes, M.O., Andrews, M.R., van Druten, N.J., Durfee, D.S., Kurn, D.M., Ketterle, W.: Bose-einstein condensation in a gas of sodium atoms. *Physical Review Letters* 75, 3969–3973 (1995)

4. Heisenberg, W.: Über den anschaulichen inhalt der quantentheoretischen kinematik und mechanik. *Zeitschrift für Physik* 43, 172–198 (1927)
5. Klempt, C., Topic, O., Gebreyesus, G., Scherer, M., Henninger, T., Hyllus, P., Ertmer, W., Santos, L., Arlt, J.J.: Multiresonant spinor dynamics in a bose-einstein condensate. *Physical Review Letters* 103, 195302 (2009)
6. Klempt, C., Topic, O., Gebreyesus, G., Scherer, M., Henninger, T., Hyllus, P., Ertmer, W., Santos, L., Arlt, J.J.: Parametric amplification of vacuum fluctuations in a spinor condensate. *Physical Review Letters* 104, 195303 (2010)
7. Scherer, M., Lücke, B., Gebreyesus, G., Topic, O., Deuretzbacher, F., Ertmer, W., Santos, L., Arlt, J.J., Klempt, C.: Spontaneous breaking of spatial and spin symmetry in spinor condensates. *Physical Review Letters* 105, 135302 (2010)
8. Castin, Y., Dum, R.: Bose-einstein condensates in time dependent traps. *Physical Review Letters* 77, 5315–5319 (1996)
9. Tuytelaars, T., Mikolajczyk, K.: Local invariant feature detectors: a survey. *Foundations and Trends in Computer Graphics and Vision*, vol. 3 (2008)
10. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. *International Journal of Computer Vision (IJCV)* 60, 63–86 (2004)
11. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust wide baseline stereo from maximally stable extremal regions. In: *British Machine Vision Conference (BMVC)*, vol. 1, pp. 384–393 (2002)
12. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Gool, L.V.: A comparison of affine region detectors. *International Journal of Computer Vision (IJCV)* 65, 43–72 (2005)
13. Lindeberg, T.: Feature detection with automatic scale selection. *International Journal of Computer Vision (IJCV)* 30, 79–116 (1998)
14. Lindeberg, T., Garding, J.: Shape-adapted smoothing in estimation of 3-d shape cues from affine deformations of local 2-d brightness structure. *Image and Vision Computing (IVC)* 15, 415–434 (1997)
15. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision (IJCV)* 60, 91–110 (2004)
16. Cordes, K., Müller, O., Rosenhahn, B., Ostermann, J.: Bivariate feature localization for sift assuming a gaussian feature shape. In: *Bebis, G., Boyle, R., Parvin, B., Koracin, D., Chung, R., Hammoud, R., Hussain, M., Kar-Han, T., Crawfis, R., Thalmann, D., Kao, D., Avila, L. (eds.) ISVC 2010. LNCS, vol. 6453, pp. 264–275. Springer, Heidelberg (2010)*

A Graph-Kernel Method for Re-identification

Luc Brun¹, Donatello Conte², Pasquale Foggia², and Mario Vento²

¹ GREYC UMR CNRS 6072
ENSICAEN-Université de Caen Basse-Normandie,
14050 Caen, France

luc.brun@greyc.ensicaen.fr

² Dipartimento di Ingegneria dell'Informazione e di Ingegneria Elettrica,
Università di Salerno, Via Ponte Don Melillo, 1 I-84084 Fisciano (SA), Italy
{dconte,pfoggia,mvento}@unisa.it

Abstract. Re-identification, that is recognizing that an object appearing in a scene is a reoccurrence of an object seen previously by the system (by the same camera or possibly by a different one) is a challenging problem in video surveillance. In this paper, the problem is addressed using a structural, graph-based representation of the objects of interest. A recently proposed graph kernel is adopted for extending to this representation the Principal Component Analysis (PCA) technique. An experimental evaluation of the method has been performed on two video sequences from the publicly available PETS2009 database.

1 Introduction

In the last years, research in the field of intelligent video surveillance has progressively shifted from low-level analysis tasks (such as object detection, shadow removal, short term tracking etc.) to high-level event detection (including long term tracking, multicamera tracking, behavior analysis etc.).

An important task required by many event detection methods is to establish a suitable correspondence between observations of people who might appear and reappear at different times and across different cameras. This kind of problematic is commonly known as “people re-identification”.

Several applications using single camera setup may benefit from information induced by people re-identification. One of the main applications is loitering detection. Loitering refers to prolonged presence of people in an area. This behaviour is interesting in order to detect, for example, beggars in street corners, or drug dealers at bus stations, and so on. Beside this, information on these re-occurrences is very important in multi-camera setups, such as the ones used for wide area surveillance. Such surveillance systems create a novel problem of discontinuous tracking of individuals across large sites, which aims to reacquire a person of interest in different non-overlapping locations over different camera views.

Re-identification problem has been studied for last five years approximately. A first group [10,18,34] deals with this problem by defining a unique signature

which condenses a set of frames of a same individual; re-identification is then performed using a similarity measure between signatures and a threshold to assign old or new labels to successive scene entrances. In [10] a panoramic map is used to encode the appearance of a person extracted from all cameras viewing it. Such a method is hence restricted to multicamera systems. The signature of a person in [18] is made by a combination of SIFT descriptors and color features. The main drawback of this approach is that people to be added into the database are manually provided by a human operator. In [3] two human signatures, which use haar-like features and dominant color descriptor (DCD) respectively, are proposed while in [4] the signature is based on three features, one capturing global chromatic information and two analyzing the presence of recurrent local patterns.

A second group ([26,5]) deals with re-identification of people by means of a representation of a person in a single frame. Each representation corresponds to a point in a feature space. Then a classification is performed by clustering these points using a SVM ([26]) or a correlation module ([5]). Both [26,5] use the so-called “color-position” histogram: the silhouette of a person is first vertically divided into n equal parts and then some color features (RGB mean, or HSV mean, etc.) are computed to characterize each part.

This paper can be ascribed to the second group but with some significant novelty: first, we have a structural (graph-based) representation of a person; second, our classification scheme is based on *graph kernels*. A graph kernel is a function in graph space that shares the properties of the dot-product operator in vector space, and so can be used to apply many vector-based algorithms to graphs.

Many graph kernels proposed in the literature have been built on the notion of *bag of patterns*. Graphlets kernels [22] are based on the number of common sub-graphs of two graphs. Vert [15] and Borgwardt [23] proposed to compare the set of sub-trees of two graphs. Furthermore, many graph kernels are based on simpler patterns such as walks [14], trails [9] or paths.

A different approach is to define a kernel on the basis of a graph edit distance, that is the set of operations with a minimal cost transforming one graph into another. Kernels based on this approach do not rely on the (often simplistic) assumption that a bag of patterns preserves most of the information of its associated graph. The main difficulty in the design of such graph kernels is that the edit distance does not usually corresponds to a metric. Trivial kernels based on edit distances are thus usually non definite positive. Neuhaus and Bunke [16] proposed several kernels based on edit distances. These kernels are either based on a combination of graph edit distances (trivial kernel, zeros graph kernel), use the convolution framework introduced by Haussler [12] (convolution kernel, local matching kernel), or incorporate within the kernel construction schemes several features deduced from the computation of the edit distance (maximum similarity edit path kernel, random walk edit kernel). Note that a noticeable exception to this classification is the diffusion kernel introduced by the same authors [16]

which defines the gram matrix associated to the kernel as the exponential of a similarity matrix deduced from the edit distance.

We propose in this paper to apply a recent graph kernel [6,11] based on edit distance, together with statistical machine learning methods, to people re-identification. The remaining of this paper is structured as follows: we first describe in Section 2 our graph encoding of objects within a video. Moving objects are acquired from different view points and are consequently encoded by a set of graphs. Given such a representation we describe in Section 3 an algorithm which allows to determine if a given input graph corresponds to a new object. If this is not the case, the graph is associated to one of the objects already seen. The different hypotheses used to design our algorithm are finally validated through several experiments in Section 4.

2 Graph-Based Object Representation

The first step of our method aims to separate pixels depicting people on the scene (foreground) from the background. We thus perform a detection of moving areas, by background subtraction, combined with a shadow elimination algorithm [7]. This first step provides a set of masks which is further processed using mathematical morphology operations (closing and opening) (Fig. 1a). Detected foreground regions are then segmented using Statistical Region Merging (SRM) algorithm [17] (Fig. 1c). Finally, the segmentation of the mask within each rectangle is encoded by a Region adjacency Graph (RAG). Two nodes of this graph are connected by an edge if the corresponding regions are adjacent. Labels of a node are: the RGB average color and the normalized size η defined as the ratio between the area of the region and the one of the overall image (Fig. 1d).

3 Comparisons between Objects by Means of Graph Kernels

Objects acquired by multiple cameras, or across a large time interval, may be subject to large variations. Common kernels [14] based on walks, trails or paths are quite sensitive to such variations. On the other hand, graph edit distances correspond to the minimal overall cost of a sequence of operations transforming two graphs. Within our framework, such distances are parametrized by two sets of functions $c(u \rightarrow v)$, $c(u \rightarrow \epsilon)$ and $c(e \rightarrow e')$, $c(e \rightarrow \epsilon)$ encoding respectively the substitution, and deletion costs for nodes and edges. Using such distances, small graph distortions may be encoded by small edit costs, hence allowing to capture graph similarities over sets having important within-class distance. Unfortunately, the computational complexity of the exact edit distance is exponential in the number of involved nodes, which drastically limits its applicability to databases composed of small graphs.

This paper is based on a sub optimal estimation of the edit distance proposed by Nehauss and Bunke [19,20]. Let us consider two labeled graphs

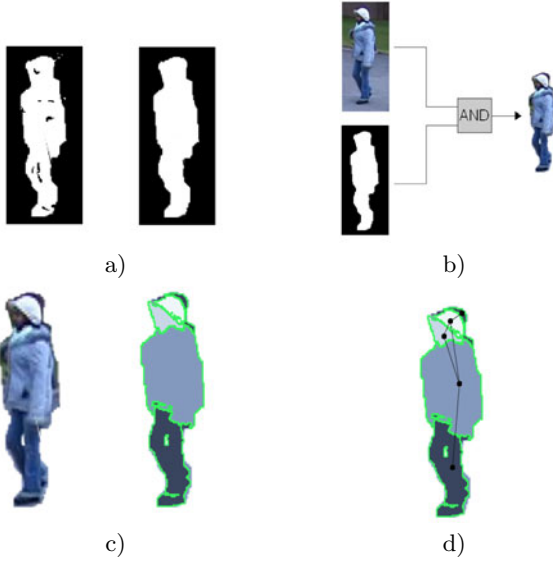


Fig. 1. a) Application of a suited morphological operator; b) Extraction of person appearance; c) Image segmentation; d) RAG construction

$g_1 = (V_1, E_1, \mu_1)$ and $g_2 = (V_2, E_2, \mu_2)$ where μ_1 and μ_2 denote respectively the vertex's labels of g_1 and g_2 . For any vertex w of V_1 or V_2 , let us further denote by $\angle(w)$ the set of edges incident to w . The distance between g_1 and g_2 is estimated by first computing for each couple of vertices $(u, v) \in V_1 \times V_2$ the best mapping between $\angle(u)$ and $\angle(v)$. Such a mapping is defined by a permutation σ from a set $S_{u,\sigma} \subset \angle(u)$ to a set $S_{v,\sigma} \subset \angle(v)$, the remaining edges $\angle(u) - S_{u,\sigma}$ and $\angle(v) - S_{v,\sigma}$ being respectively denoted by $N_{u,\sigma}$ and $N_{v,\sigma}$. The cost of a mapping is defined as the overall cost of edges substitutions from $S_{u,\sigma}$ to $S_{v,\sigma}$ and edge deletions from $N_{u,\sigma}$ and $N_{v,\sigma}$. The optimal mapping, denoted $\Delta^e(u, v)$ being defined as the mapping of minimal cost:

$$\Delta^e(u, v) = \min_{\sigma \in M_{u,v}} \sum_{e \in S_{u,\sigma}} c(e \rightarrow \sigma(e)) + \sum_{e \in N_{u,\sigma} \cup N_{v,\sigma}} c(e \rightarrow \epsilon)$$

where $M_{u,v}$ denotes the set of mappings from $\angle(u)$ to $\angle(v)$.

This optimal mapping is determined using the Hungarian Algorithm [20] applied on the sets $\angle(u)$ and $\angle(v)$. The total cost of mapping vertex u to vertex v together with the sets of incident edges of both vertices is denoted $\Delta(u, v) = c(u \rightarrow v) + \Delta^e(u, v)$. The Hungarian algorithm between V_1 and V_2 based on the cost functions $\Delta(u, v)$ and $c(u \rightarrow \epsilon)$ provides an optimal mapping σ^* between the nodes of both sets denoted $Editcost(g_1, g_2)$:

$$Editcost(g_1, g_2) = \sum_{u \in S_{1,\sigma^*}} \Delta(u, \sigma^*(u)) + \sum_{u \in N_{1,\sigma^*} \cup N_{2,\sigma^*}} c(u \rightarrow \epsilon) \quad (1)$$

where S_{1,σ^*} (resp. S_{2,σ^*}) corresponds to the set of vertices of V_1 (resp. V_2) mapped to some vertices of V_2 (resp. V_1) by the optimal mapping σ^* while N_{1,σ^*} (resp. N_{2,σ^*}) corresponds to the set of deleted vertices in g_1 (resp. g_2).

Now we will discuss the four cost functions used for defining the edit distance. Within our framework, each node u encodes a region and is associated to the mean color (R_u, G_u, B_u) and to the normalized size η_u of the region (Section 2). We experimentally observed that small regions have larger chances to be deleted between two segmentations. Hence, the normalized size of a region can be used as a measure of its relevance within the whole graph.

The cost of a node substitution is defined as the distance between the mean colors of the corresponding regions. We additionally weigh this cost by the maximum normalized size of both nodes. Such a weight avoids to penalize the matching of small regions, which should have a small contribution to the global similarity of both graphs. Also, a term is added to account for the size difference between the regions:

$$c(u \rightarrow v) = \max(\eta_u, \eta_v) \cdot d_c(u, v) + \gamma_{NodeSize} \cdot |\eta_u - \eta_v| \quad (2)$$

where $d_c(u, v)$ is the distance in the color space, and $\gamma_{NodeSize}$ is a weight parameter selected by cross validation. The distance $d_c(u, v)$ is not computed as the Euclidean distance between RGB vectors, but uses the following definition that is based on the human perception of colors [1]:

$$d_c(u, v) = \sqrt{\left(2 + \frac{\bar{\tau}}{2^k}\right)\delta_R^2 + 4\delta_G^2 + \left(2 - \frac{(2^k - 1) - \bar{\tau}}{2^k}\right)\delta_B^2} \quad (3)$$

where k is the channel depth of the image, $\bar{\tau} = \frac{R_u + R_v}{2}$ and δ_R, δ_G and δ_B encode respectively the differences of coordinates along the red, green and blue axis.

The cost of a node deletion should be proportional to its relevance encoded by the normalized size, and is thus defined as:

$$c(u \rightarrow \epsilon) = \gamma_{NodeSize} \cdot \eta_u \quad (4)$$

Using the same basic idea, the cost of an edge removal should be proportional to the minimal normalized size of its two incident nodes.

$$c((u, u') \rightarrow \epsilon) = \gamma_{Edge} \cdot \gamma_{EdgeSize} \cdot \min(\eta_u, \eta_{u'}) \quad (5)$$

where $\gamma_{EdgeSize}$ encodes the specific weight of the edge removal operation while γ_{Edge} corresponds to a global edge's weight.

Equation 4 and 5 are based on the implicit assumption that the main variations between two segmentations are induced by small regions which may appear or disappear between two successive segmentations. Note that one may additionally consider the possibility that two adjacent regions may be merged in one segmentation and not in the other. Taking into account such phenomena may improve our node and edge deletion cost at the price of additional parameters within equations 4 and 5.

Within a region adjacency graph, edges only encode the existence of some common boundary between two regions. Moreover, these boundaries may be drastically modified between two segmentations. Therefore, we choose to base the cost of an edge substitution solely on the substitution's cost of its two incident nodes.

$$c((u, u') \rightarrow (v, v')) = \gamma_{Edge} \cdot (c(u \rightarrow v) + c(u' \rightarrow v')) \quad (6)$$

Note that all edge costs are proportional to the weight γ_{Edge} . This last parameter allows thus to balance the importance of node and edge costs.

3.1 From Graph Edit Distance to Graph Kernels

Let us consider a set of input graphs $\{G_1, \dots, G_n\}$ defining our graph test database. Our person re-identification is based on a distance of an input graph G from the space spanned by $\{G_1, \dots, G_n\}$. Such a measure of novelty detection requires to embed the graphs into a metric space. Given our edit distance (Section 3), one may build a $n \times n$ similarity matrix $W_{i,j} = \exp(-EditCost(G_i, G_j)/\sigma)$ where σ is a tuning variable. Unfortunately, the edit distance does not fulfill all the requirements of a metric; consequently, the matrix W may be not semi-definite and hence does not define a kernel.

As mentioned in Section 1, several kernels based on the edit distance have been recently proposed. However, these kernels are rather designed to obtain a definite positive matrix of similarity than to explicitly solve the problem of kernel-based classification or regression methods. We thus use a recent kernel construction scheme [6,11] based on an original remark by Steinke [24]. This scheme [6,11] exploits the fact that the inverse of any regularised Laplacian matrix deduced from W defines a definite positive matrix and hence a kernel on $\{G_1, \dots, G_n\}$. Thus, our kernel construction scheme first builds a regularised Laplacian operator $\tilde{L} = I + \lambda L$, where λ is a regularisation coefficient and L denotes the normalized Laplacian defined by: $L = I - D^{-\frac{1}{2}} W D^{-\frac{1}{2}}$ and D is a diagonal matrix defined by $D_{i,i} = \sum_{j=1}^n W_{i,j}$. Our kernel is then defined as: $K = \tilde{L}^{-1}$. Using a classification or regression scheme, such a kernel leads to map graphs having a small edit distance [6,11] (and thus a strong similarity) to close values.

3.2 Novelty Detection and Person Re-identification

Within our framework, each reappeared person is represented by a set of graphs encoding the different acquisitions of this person. Before assigning a new input graph to an already created class, we must determine if this graph corresponds to a person already encountered. This is a problem of novelty detection, with the specific constraint that each class of graphs encoding an already encountered person has a large within-class variation. Several methods, such as one class SVM [21] or support vector domain description [25] have been used for novelty detection. However, these methods are mainly designed to compare an incoming data with an homogeneous data set. The method of Desobry [8] has the same

drawback and is additionally mainly designed to compare two sets rather than one set with an incoming datum.

The method introduced by Hoffman [13] is based on kernel Principal Component Analysis (PCA). An input datum is considered as non belonging to a class if its squared distance from the space spanned by the first principal components of the class is above a given threshold. Note that this method is particularly efficient using high dimensional spaces such as the one usually associated to kernels. This method has the additional advantage of not assuming a strong homogeneity of the class.

Given an input graph G and a set of k classes, our algorithm first computes the set $\{d_1(G), \dots, d_k(G)\}$ where $d_i(G)$ is the squared distance of the input graph G from the space spanned by the first q principal component of class i . Our novelty decision criterion is then based on a comparison of $d(G) = \min_{k=1,n} d_k(G)$ against a threshold.

If $d(G)$ is greater than the specified threshold, G is considered as a new person entering the scene. Otherwise, G describes an already encountered person, which is assigned to the class i that minimizes the value of $d_i(G)$.

4 Experimental Results

We implemented the proposed method in C++ and tested its performance on two video sequences taken from the PETS2009 [2] database (Fig. 2). Each video sequence is divided in two parts so as to build the training and test sets. In this experiment we have used one frame every 2 seconds from each video, in order to have different segmentations of each person. The training set of the first sequence (View001) is composed of 180 graphs divided into 8 classes, while the test set contains 172 graphs (30 new and 142 existing). The second sequence (View005) is composed of 270 graphs divided into 9 classes for the training set, and 281 graphs (54 new and 227 existing) for the test set.

In order to evaluate the performances of the algorithm, we have used the following measures:

- The **true positives rate (TP)**, i.e the rate of test patterns correctly classified as novel (positive): $TP = \text{true positive}/\text{total positive}$



a) View 001



b) View 005

Fig. 2. Sample frames from the PETS2009 dataset

- The **false positives rate (FP)**, i.e the rate of test patterns incorrectly classified as novel (positive): $TP = \text{false positive}/\text{total negative}$
- The **detection accuracy (DA)**:

$$DA = (\text{true positive} + \text{true negative})/(\text{total positive} + \text{total negative})$$

- The **classification accuracy (CA)**, i.e the rate of samples classified as negatives which are then correctly classified with multi-class SVM
- The **Total Accuracy**: $TA = DA \times CA$.

As shown on Fig. 3 we obtained around 85% of novelty detection accuracy, and 70% of total accuracy for both View001 and View005 sequences. These results were obtained with the Graph Laplacian Kernel using $\sigma = 4.7$ and $\lambda = 10.0$.

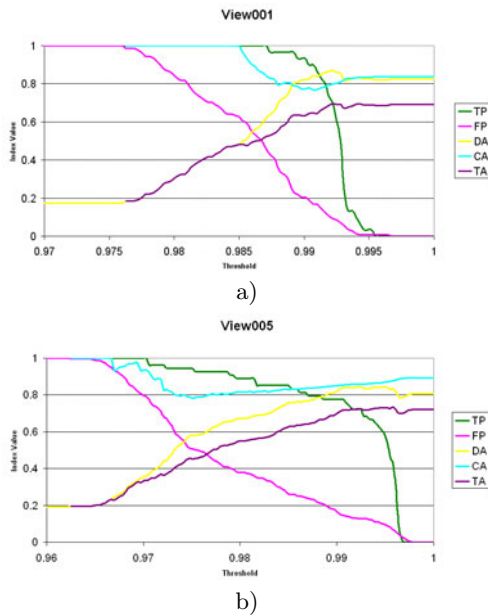


Fig. 3. Performances result on the view001 (a) and view005 (b) of the PETS2009 dataset

These results appear very promising. For a wide interval of threshold values the classification accuracy rate remains close to 100%. Furthermore, the True Positive Rate curve has a high slope in correspondence of a high value of the threshold, while the False Positive Rate has a smoother behavior; this means that the algorithm can reliably find a threshold value that is able to discard most of the false positives while keeping most of the true positives. Finally, the ROC curves (Fig. 4) are close to the upper and left edges of the True Positive/False Positive space, confirming the discriminant power of the proposed method.

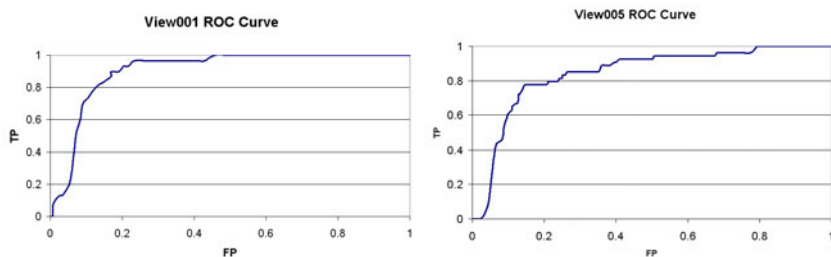


Fig. 4. ROC curves for the two sequences from the PETS2009 dataset

5 Conclusions

This paper presents a novel method for people re-identification based on a graph-based representation and a graph kernel. It combines our graph kernel with a novelty detection method based on Principal Component Analysis in order to detect if an incoming graph corresponds to a new person and, if not, to correctly assign the identity of a previously seen person. Our future works will also extend the present method to people re-identification within groups. In such cases, a whole group is encoded by a single graph. Thus, the used kernel should be able to match subgraphs within larger graphs. We plan to study the ability of graphlet kernels to perform this task.

References

1. <http://www.compuphase.com/cmetric.htm>
2. Database: Pets 2009 (2009), <http://www.cvg.rdg.ac.uk/PETS2009/>
3. Bak, S., Corvee, E., Brmond, F., Thonnat, M.: Person re-identification using haar-based and dcd-based signature. In: 2010 Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (2010)
4. Bazzani, L., Cristani, M., Perina, A., Farenzena, M., Murino, V.: Multiple-shot person re-identification by hpe signature. In: Proceedings of 20th International Conference on Pattern Recognition, ICPR 2010 (2010)
5. Bird, N., Masoud, O., Papanikolopoulos, N., Isaacs, A.: Detection of loitering individuals in public transportation areas. *IEEE Transactions on Intelligent Transportation Systems* 6(2), 167–177 (2005)
6. Brun, L., Conte, D., Foggia, P., Vento, M., Villemin, D.: Symbolic learning vs. graph kernels: An experimental comparison in a chemical application. In: 14th Conf. on Advances in Databases and Information Systems (ADBIS) (2010)
7. Conte, D., Foggia, P., Percannella, G., Vento, M.: Performance evaluation of a people tracking system on pets2009 database. In: Seventh IEEE International Conference on Advanced Video and Signal Based Surveillance (2010)
8. Desobry, F., Davy, M., Doncarli, C.: An online kernel change detection algorithm. *IEEE Transaction on Signal Processing* 53(8), 2961–2974 (2005)
9. Dupé, F.X., Brun, L.: Tree covering within a graph kernel framework for shape classification. In: Foggia, P., Sansone, C., Vento, M. (eds.) *ICIAP 2009*. LNCS, vol. 5716, pp. 278–287. Springer, Heidelberg (2009)

10. Gandhi, T., Trivedi, M.M.: Panoramic appearance map (pam) for multi-camera based person re-identification. In: IEEE International Conference on Video and Signal Based Surveillance, AVSS 2006 (2006)
11. Gauzere, B., Brun, L., Villemin, D.: Graph edit distance and treelet kernels for chemoinformatic. In: Graph Based Representation 2011, IAPR-TC15, Munster, Germany (May 2011) (submitted)
12. Haussler, D.: Convolution kernels on discrete structures. Tech. rep., Department of Computer Science, University of California at Santa Cruz (1999)
13. Hoffmann, H.: Kernel pca for novelty detection. *Pattern Recognition* 40(3), 863 (2007)
14. Kashima, H., Tsuda, K., Inokuchi, A.: Marginalized kernel between labeled graphs. In: Proc. of the Twentieth International conference on Machine Learning (2003)
15. Mah, P., Vert, J.P.: Graph kernels based on tree patterns for molecules. *Machine Learning* 75(1), 3–35 (2008)
16. Neuhaus, M., Bunke, H.: Bridging the Gap Between Graph Edit Distance and Kernel Machines. World Scientific Publishing Co., Inc., River Edge (2007)
17. Nock, R., Nielsen, F.: Statistical region merging. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 26(11), 1452–1458 (2004)
18. de Oliveira, I.O., de Souza Pio, J.L.: People reidentification in a camera network. In: IEEE Int. Conf. on Dependable, Autonomic and Secure Computing (2009)
19. Riesen, K., Bunke, H.: Approximate graph edit distance computation by means of bipartite graph matching. *Image Vision Computing* 27(7), 950–959 (2009)
20. Riesen, K., Neuhaus, M., Bunke, H.: Bipartite graph matching for computing the edit distance of graphs. In: Escolano, F., Vento, M. (eds.) GBRPR. LNCS, vol. 4538, pp. 1–12. Springer, Heidelberg (2007)
21. Scholkopf, B., Platt, J., Shawe-Taylor, J., Smola, A.J., Williamson, R.C.: Estimating the support of a high-dimensional distribution. *Neural Computation* 13, 1443–1471 (2001)
22. Shervashidze, N., Vishwanathan, S.V., Petri, T.H., Mehlhorn, K., Borgwardt, K.M.: Efficient graphlet kernels for large graph comparison. In: Twelfth International Conference on Artificial Intelligence and Statistics (2009)
23. Shervashidze, N., Borgwardt, K.: Fast subtree kernels on graphs. In: Advances in Neural Information Processing Systems 22. Curran Associates Inc. (2009)
24. Steinke, F., Schököpf, B.: Kernels, regularization and differential equations. *Pattern Recognition* 41(11), 3271–3286 (2008)
25. Tax, D., Duin, R.: Support vector domain description. *Pattern Recognition Letters* 20, 1191–1199 (1999)
26. TruongCong, D.N., Khoudour, L., Achard, C., Meurie, C., Lezoray, O.: People re-identification by spectral classification of silhouettes. *Signal Processing* 90, 2362–2374 (2010)

Automatic Recognition of 2D Shapes from a Set of Points

Benoît Presles¹, Johan Debayle¹, Yvan Maillot², and Jean-Charles Pinoli¹

¹ CIS-SPIN-LPMG / CNRS

École Nationale Supérieure des Mines de Saint-Étienne,
158 cours Fauriel F-42023 Saint-Étienne cedex 2, France
`presles@emse.fr`

² Laboratoire MIA, Équipe MAGE, Université de Haute-Alsace,
4 rue des frères Lumière, 68093 Mulhouse cedex, France

Abstract. 2D shape recognition from a set of points is largely used in several imaging areas such as geometric modeling, image visualization or medical image analysis. However, the perceived shape of a set of points is subjective. It is mainly influenced by the spatial arrangement of the points and by several cognitive factors. The Delaunay filtration methods derived from the well-known α -shapes, like LDA- α -shapes or conformal- α -shapes, provide a family of shapes capturing the intuitive notion of “crude” versus “fine” shape of a set of points. In this paper, a quantitative criterion based on shape measurements is defined for extracting the “optimal” shape from this family that best corresponds to the human visual perception. A novel automatic shape recognition method is proposed and successfully evaluated on the KIMIA image database, where the reference shapes are known and sampled by generating 2D point sets.

Keywords: Convexity, Delaunay triangulation, Human visual perception, LDA- α -shapes, Pattern recognition.

1 Introduction

The shape of a set of points may be quite naturally perceived from a human point of view but it is rather difficult to be calculated by a computer. Indeed, a lot of shapes can characterize a cloud of points. Its convex hull is one of them, but it is most of the time relatively far from those one humanly perceived. So, a mathematical concept to apprehend the shape of a set of points might be helpful in order to find an appropriate shape by computing.

Jarvis in [5] was the first to consider the shape as a generalization of the convex hull. A few years later, Edelsbrunner, Kirkpatrick and Seidel in [3], gave a very powerful theory for reasoning about the shape of a set of points. This seminal concept are the well-known α -shapes, a formal definition of a generalized convex hull. The α -shapes are a filtration of the Delaunay triangulation depending on α , a real number. They use a distance between two connected points of the Delaunay triangulation to decide which edges belong to an α -shape. The α -shapes lead to a discrete family of shapes, from the “crudest” to the “finest”

shape. The “crudest shape” is the whole plane when $\alpha = \infty$. The next shape is the polygon inscribed in the smallest circle enclosing all the points and whose vertices are the ones belonging to the boundary of the circle. An other shape is the convex hull when $\alpha = 0$. And the “finest” shape is the set of points itself when $\alpha = -\infty$. There is a lot of non-convex shapes between 0 and $-\infty$. It is reasonable to think that the “desired” shape is close to one of them if the points are uniformly distributed and if the set of points is dense enough.

However, there are situations where significant shapes do not belong to this family, for example when the point set is not uniformly distributed. In order to overcome this difficulty, a few of related notions were introduced, among them:

- The *weighted- α -shapes* [4] were the first response of Edelsbrunner. This is the α -shapes with weighted points, where large weights can be assigned in sparse region and small weights in dense region in order to counteract the problem of a non-uniform distribution. But the way to compute the weights is difficult and sometimes impossible.
- The *\mathcal{A} -shapes* [8] are also a sub-graph of the Delaunay triangulation. Moreover, it contains the α -shapes. But it depends on a parameter \mathcal{A} which is a point set, and there is no really efficient algorithm to compute it.
- The *conformal- α -shapes* were introduced in [1]. This filtration depends on two local parameters, α_p^- and α_p^+ for each point p , and on a global variable α that can vary from 0 to ∞ . Both parameters α_p^- and α_p^+ are computed according to the neighborhood of p so that the variation of α leads to a filtration depending on a local scale parameter fixed by α_p^- and α_p^+ .
- The *LDA- α -shapes* were presented in [6]. This Delaunay filtration depends on one variable α , a real number in $[0, 1]$. It takes into account the local density of the points in order to admit non-uniform distributions for the point set.

Note that, α -shapes, conformal- α -shapes, and LDA- α -shapes all produce a Delaunay filtration and can be generalized in any spatial dimension. In the present work, only planar shapes and more precisely non-convex polygons enclosing the whole point set are investigated.

The aim of this paper is to present a method that automatically find the “good” shape from such kind of Delaunay filtrations according to the human visual perception (Gestalt laws). Now the problem is to find, among a large set of filtration methods, the one that fits our aim. To do our experiments, we have chosen the LDA- α -shapes because:

1. This method considers non-uniform point sets (more suitable with our specific problem).
2. It depends on one variable only.
3. Its algorithm is very simple, efficient and easy to implemented.

But this could be tested with other Delaunay filtrations like the α -shapes or the conformal- α -shapes. Note that it should not be possible for the \mathcal{A} -shapes.

2 LDA- α -Shapes

The aim of the LDA- α -shapes (short for Locally-Density-Adaptive- α -shapes) is to reconstruct a domain from a “well-distributed” point set, even if it is not uniformly distributed. So, the point set may be more or less dense in places according to the local required amount of details. As shown on Figure 1, the density of the point set is strong close to the eyes of the lizard or on its fingers, while it is sparse on its back and on its tail. Since, on the one hand, a sudden variation of the local density indicates the presence of a hole or a hollow, on the other hand, variations of the local density are allowed, the point density must change gradually to avoid the formation of non-existent holes. These observations



Fig. 1. A shape to reconstruct (left) and a possible sample (right)

lead to the definition of the LDA- α -shapes. The Delaunay triangulation can be efficiently used to measure the density variation of a point set. For instance, in Figure 2, the disks D and D' circumscribed to the (white) Delaunay triangles are much larger than some of their neighborhood “circumdisks”. This means that there is a wide area with no point inside $D \cup D'$ (by definition) surrounded by a denser area. Therefore, there is probably a hole at this place, and, in that case, the edge (dashed) shared by both white triangles may be eliminated. More precisely, it is removed if the ratio between a disk and some of its neighborhoods is more than $1/\alpha$. Note that the LDA- α -shapes are indeed very close to the conformal- α -shapes [1] with α_p^- equals 0 and α_p^+ equals the radius of the smallest circumcircle to a Delaunay triangle whose p is a vertex. The variation of α from 0

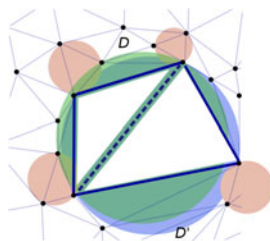


Fig. 2. There is probably a hole

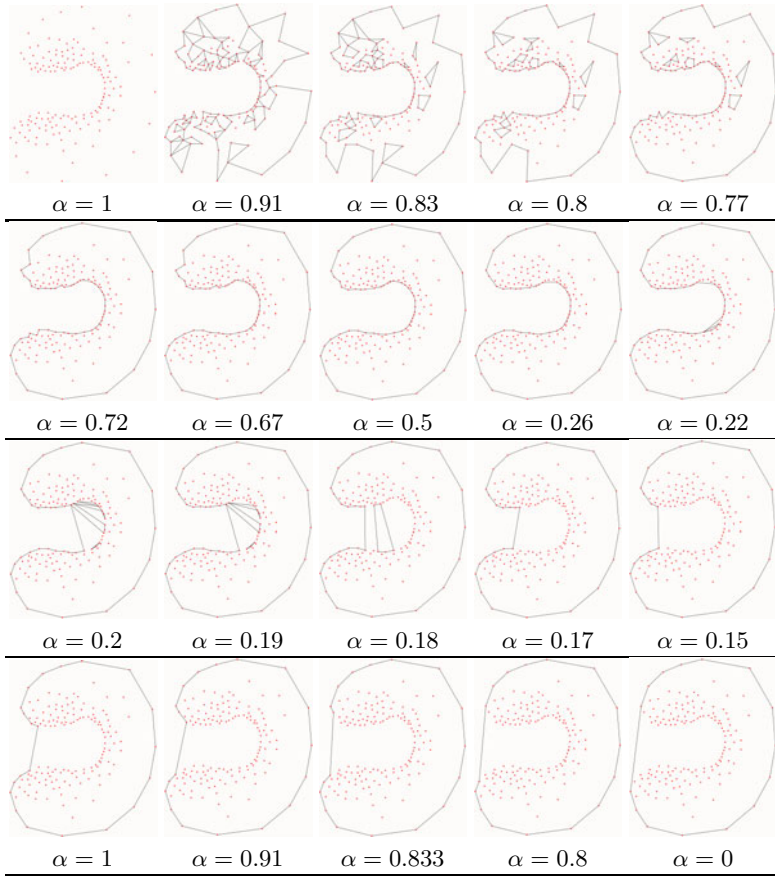


Fig. 3. A few LDA- α -shapes of the same sample according to different values of α

to 1 leads the LDA- α -shapes to form an ordered discrete family of straight-line graphs, the LDA-0-shape of a set of points being its convex hull, the LDA-1-shape being the set of points itself. Figures 3 and 4 show such a variation with a few of LDA- α -shapes of the same point set with different values of α .

The issue now is to find the appropriate shape, that is to say the “optimal” α value which enables the best approximation of the perceived shape, among a wide range of candidate shapes according to the human visual perception. For example, in Figure 3, it would be one of them between the LDA-0.22-shape and the LDA-0.72-shape or around LDA-0.5-shape in Figure 4. Note that to succeed, the expected shape has to belong to the “family” of LDA- α -shapes. If the α -shapes had been chosen instead of the LDA- α -shapes, with the same sample as in Figures 3 or 4, it would not be able to find an appropriate shape, whatever α , because of the sparsity to the right of the point set in Figure 3 or on the back of the lizard in Figure 4.

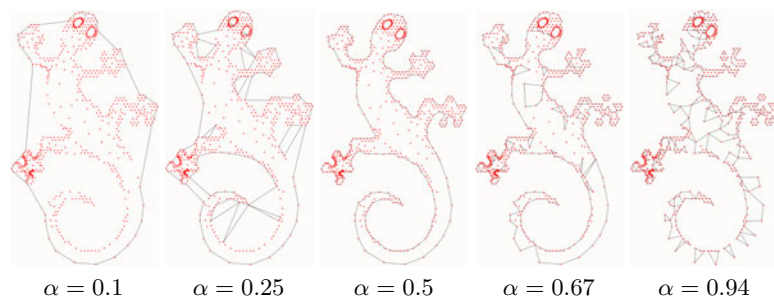


Fig. 4. A few LDA- α -shapes of the same sample according to different values of α

3 Automatic Recognition of the Perceived Shape

In the literature, a very few papers [7] [2] have investigated this automatic parameter selection in relation with the expected shape. In both papers, this selection is based on the concept of minimum spanning trees. The second paper also defines a criterion based on the the edge lengths of the Delaunay triangulation. Nevertheless, in both papers, the selection of the optimal parameter is proposed without any justification in relation with the human visual perception. In this section, a quantitative evaluation of the perceived shape from a set of points is proposed so as to automatically pick the optimal α value consistently with the human visual perception.

3.1 What is the Perceived Shape of a Set of Points?

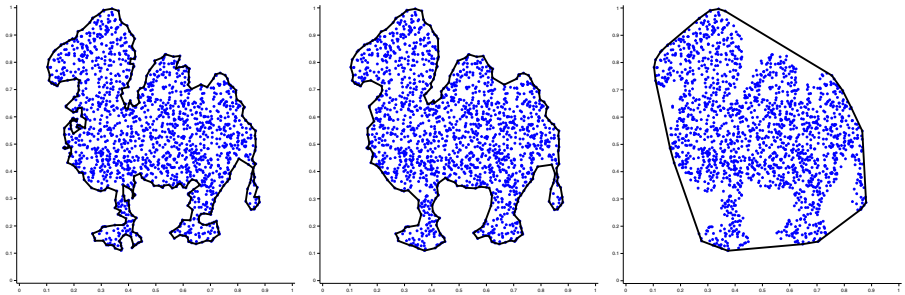
It is assumed in this paper that the studied set of points is noise free and represents only a simply connected compact set. Consequently, only the exterior hull of the LDA- α -shape is taken into account (the shape has no holes). For example in Figure 3 for $\alpha = 0.77$, the edges which belong to the LDA- α -shape inside the exterior hull are not considered. Under these assumptions, the investigation can be restricted to LDA- α -shapes having specific properties:

- All the points should be interior to the exterior hull of the LDA- α -shape.
- The LDA- α -shape is composed of only one connected component.

The previous properties are very easy to verify for each LDA- α -shape. In this way, only a subset $\{\alpha_1, \dots, \alpha_n = 0\}$ with $(\alpha_1 > \alpha_2 > \dots > \alpha_n = 0)$ of $\{\alpha_k\}_{k \in [1, m]}$ are candidates. These restrictions are imposed in this paper but the proposed method could be extended to more general cases.

The perceived shape of a set of points is highly subjective. It is mainly influenced by the spatial arrangement of the points and by several cognitive factors. In addition, the solution may not be unique. From a visual point of view, the perceived shape should be close to the set of points while being regular (a tortuous hull is not desirable). These two characteristics can be mathematically

described: the perceived shape has minimal area and minimal perimeter. Nevertheless, these objectives are controversial. Indeed, the perimeter can only be minimized by maximizing the area. The convex hull ($\alpha_n = 0$) has maximum area and minimum perimeter while the first hull (α_1) has minimum area and maximum perimeter (Figure 5). Consequently, a “reasonable” hull achieves a compromise between reducing the area and increasing the perimeter.



(a) LDA- α_1 -shape (tortuous hull). $A=0.34, P=5.73$. (b) LDA- α_i -shape (reasonable hull). $A=0.36, P=4.41$. (c) LDA- α_n -shape (convex hull). $A=0.50, P=2.63$.

Fig. 5. Area and perimeter of different LDA- α -shapes. In black, the retrieved hull

3.2 The Proposed Criterion

First of all, to get homogeneous values between the area and the perimeter of each LDA- α -shape, denoted $A(\alpha)$ and $P(\alpha)$ respectively, it is necessary to normalize them. Let $\overline{A(\alpha)}$ and $\overline{P(\alpha)}$ be the normalized area and the normalized perimeter of each LDA- α -shape respectively:

$$\overline{A(\alpha)} = \frac{A(\alpha) - A(\alpha_1)}{A(\alpha_n = 0) - A(\alpha_1)}, \overline{P(\alpha)} = \frac{P(\alpha) - P(\alpha_n)}{P(\alpha_1) - P(\alpha_n)} \tag{1}$$

The proposed criterion to be minimized to retrieve a “reasonable” hull, depending on the α value, is defined by the L_2 -norm of the vector $(\overline{A(\alpha)}, \overline{P(\alpha)})$. Nevertheless, this norm has to be weighted according to the convexity of the expected shape. Indeed, if the perceived shape is convex, it is sufficient to only minimize the perimeter. On the contrary, if the perceived shape is tortuous holding several concavities, only the area has to be minimized. In this way, the proposed criterion, denoted $APC(\alpha)$, is defined as:

$$APC(\alpha) = \|((1 - C^6)\overline{A(\alpha)}, C^6\overline{P(\alpha)})\|_2 = \sqrt{((1 - C^6)\overline{A(\alpha)})^2 + (C^6\overline{P(\alpha)})^2} \tag{2}$$

where $C \in [0, 1]$ denotes the convexity ratio of the expected shape, calculated as:

$$C = \frac{1}{n} \sum_{i=1}^n A(\alpha_i) / A(\alpha_n) \tag{3}$$

Concerning the choice of the power 6 for the convexity ratio, it has been empirically selected after several tests. Finally, the optimal α value, denoted α_{opt} , is given by minimizing $APC(\alpha)$:

$$\alpha_{opt} = \arg \min_{\alpha} APC(\alpha) \quad (4)$$

3.3 Relevance of the Criterion

To evaluate the relevance of the proposed criterion, some experiments have been performed on the KIMIA binary image database [9], each binary image representing a discretized simply connected compact set of a certain convexity. For each image I , the boundary points of the simply connected compact set were extracted and normalized into the unit square $[0, 1] \times [0, 1]$. From these contour pixels, the polygonal hull, which corresponds to the reference shape S , was retrieved. This reference shape was then sampled with random points using either a uniform law or a specific law (namely, heterogeneous law) where the probability is higher near the medial axis of the complementary of the shape. Actually, this heterogeneous sampling corresponds to a more natural sampling for preserving shape details. In addition, different point densities (low=1000 points per unit area, medium=4500 points per unit area and high=20000 points per unit area) were tested. Figure 6 shows some examples of shape sampling.

From these samples, the optimal α_{opt} value obtained using the proposed method was computed and this value was compared with the expected $\tilde{\alpha}$ value that minimizes the area of the symmetric difference ASD between the LDA- α -shape and the reference shape S :

$$\tilde{\alpha} = \arg \min_{\alpha} ASD(\text{LDA-}\alpha\text{-shape}, S) \quad (5)$$

Hence, LDA- $\tilde{\alpha}$ -shape is the best approximation of the shape S , in the sense of the symmetric difference distance which is naturally consistent with the human visual perception. Finally, the error E between the two shapes LDA- α_{opt} -shape and LDA- $\tilde{\alpha}$ -shape was measured by calculating the normalized area of the symmetric difference:

$$E = \frac{ASD(\text{LDA-}\alpha_{opt}\text{-shape}, \text{LDA-}\tilde{\alpha}\text{-shape})}{A(\tilde{\alpha})} \times 100 \quad (6)$$

Table 1 synthesizes the results of the errors related to the proposed criterion using three images of the KIMIA database (I_1, I_2, I_3) which have different convexity ratios. For each sampling, 1000 simulations were performed to get robust statistics of the error E .

These results show the relevance of the proposed criterion. Indeed, the automatically selected LDA- α_{opt} -shape is closed to the LDA- $\tilde{\alpha}$ -shape, where the LDA- $\tilde{\alpha}$ -shape best approximates the reference shape S in the sense of the symmetric difference distance. Higher the point density is, lower the mean error μ is. For a high sampling, the mean error μ is less than 4 percent. Note that the method gives better results for the heterogeneous law than the uniform one.

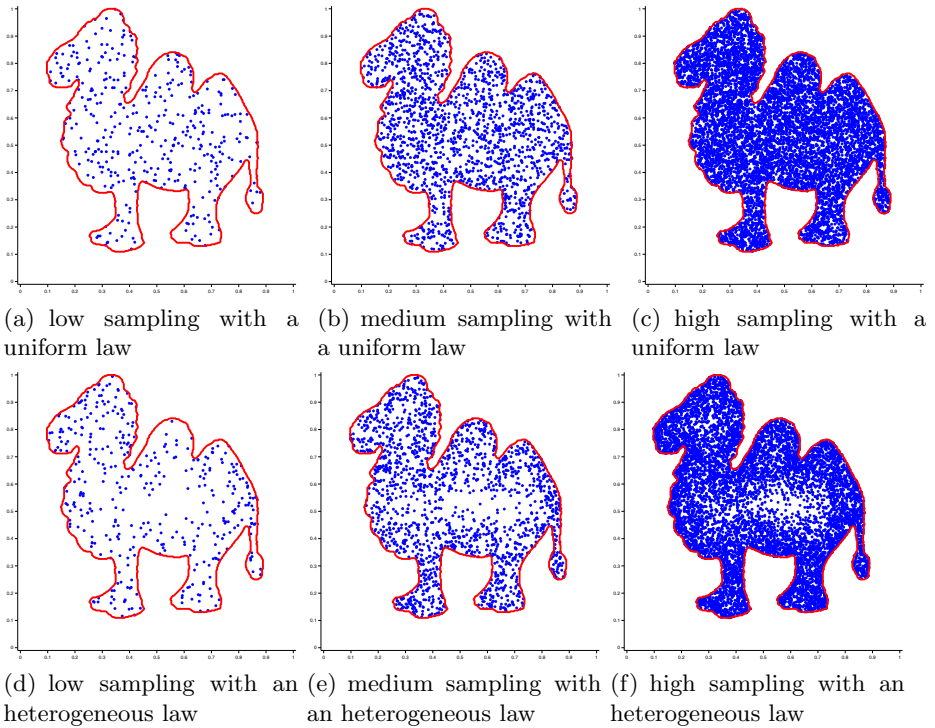


Fig. 6. Shape sampling according to three point densities (low, medium, high) and two probability laws (uniform, heterogeneous). The reference shape S is in red.

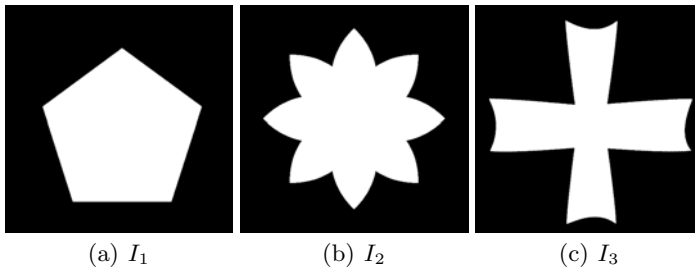


Fig. 7. Three KIMIA shape examples with different convexity ratios

4 Experiments with Quantitative Evaluation

To evaluate quantitatively the proposed method, a family of 1400 binary images from the KIMIA shape database was considered [9], each binary image representing a discretized simply connected compact set of a certain convexity. As explained in subsection 3.3, for each image, the reference polygonal hull S of the contour pixels was retrieved and the shape was sampled with random points

Table 1. Evaluation of the proposed criterion on three images with different convexity ratios. The mean error μ is in percent and is calculated according to three point densities (low, medium, high) and two probability laws (uniform, heterogeneous). σ is the standard deviation.

	uniform random sampling			heterogeneous random sampling		
	low	medium	high	low	medium	high
$E(I_1)$ ($C = 0.9$)	$\mu = 8.47$ $\sigma = 4.55$	$\mu = 2.65$ $\sigma = 0.67$	$\mu = 1.03$ $\sigma = 0.17$	$\mu = 7.62$ $\sigma = 5.32$	$\mu = 2.22$ $\sigma = 0.53$	$\mu = 0.87$ $\sigma = 0.14$
$E(I_2)$ ($C = 0.7$)	$\mu = 13.57$ $\sigma = 8.02$	$\mu = 3.61$ $\sigma = 2.45$	$\mu = 0.59$ $\sigma = 0.46$	$\mu = 12.21$ $\sigma = 8.89$	$\mu = 2.55$ $\sigma = 1.79$	$\mu = 0.48$ $\sigma = 0.39$
$E(I_3)$ ($C = 0.5$)	$\mu = 23.75$ $\sigma = 9.30$	$\mu = 10.12$ $\sigma = 4.44$	$\mu = 3.38$ $\sigma = 1.29$	$\mu = 22.79$ $\sigma = 10.61$	$\mu = 8.21$ $\sigma = 4.09$	$\mu = 2.51$ $\sigma = 0.98$

using either a uniform law or an heterogeneous law with different point densities (low, medium and high). Then, the shape boundary from the previous generated points was retrieved computing the LDA- α_{opt} -shapes. Finally, the error between the retrieve boundary and the reference boundary was quantified by calculating the normalized area of the symmetric difference between the reference shape S and the retrieved boundary.

$$E = \frac{ASD(\text{LDA-}\alpha_{opt}\text{-shape}, S)}{A(S)} \times 100 \quad (7)$$

Table 2 synthesizes the results for the 1400 binary images. Considering the het-

Table 2. Quantitative evaluation of the proposed method on the KIMIA database. The mean error μ is in percent and is calculated according to three point densities (low, medium, high) and two probability laws (uniform, heterogeneous). σ is the standard deviation.

	uniform random sampling			heterogeneous random sampling		
	low	medium	high	low	medium	high
<i>KIMIA</i> <i>database</i>	$\mu = 32.89$ $\sigma = 21.94$	$\mu = 15.90$ $\sigma = 14.15$	$\mu = 6.92$ $\sigma = 7.33$	$\mu = 31.33$ $\sigma = 22.57$	$\mu = 14.30$ $\sigma = 15.61$	$\mu = 5.95$ $\sigma = 6.72$

erogeneity of the studied shapes, the proposed method gives satisfying results. Indeed, for a high sampling, the proposed method is capable to retrieve the reference shape with a mean error μ of less than 8 percent. Note that these error values should be minimized because the error was calculating by comparing the retrieve shape with the reference shape, which can only be retrieved with an infinite density. At last, for a low sampling, the error seems quite high but actually, for such a density, the shape is poorly sampled: even visually it is very difficult to retrieve the shape boundaries. In addition, the method gives better results for an heterogeneous random sampling.

5 Conclusion and Prospects

In this paper, a novel method has been defined for automatically recognizing the “optimal” shape from a 2D point set that best corresponds to the human visual perception. It is based on the LDA- α -shapes and on a quantitative criterion defined from shape geometrical measurements (area, perimeter, convexity). The method is fully automatic without any required parameter. The performance of the proposed method has been successfully evaluated on the KIMIA image database using different point densities (low, medium, high) and sampling laws (uniform, heterogeneous). On an Intel 2.83GHz CPU running Windows XP 32 bits, the proposed algorithm needs about 1 second, 2.5 seconds and 15 seconds to retrieve α_{opt} for the low, medium and high point densities respectively. Currently, the authors try to generalize this shape recognition method to 3D point sets.

References

1. Cazals, F., Giesen, J., Pauly, M., Zomorodian, A.: Conformal alpha shapes. In: Proceedings Eurographics/IEEE VGTC Symposium Point-Based Graphics, pp. 55–61 (2005)
2. Duckham, M., Kulikb, L., Worboysc, M., Galton, A.: Efficient generation of simple polygons for characterizing the shape of a set of points in the plane. *Pattern Recognition* 41(10), 3224–3236 (2008)
3. Edelsbrunner, H., Kirkpatrick, D., Seidel, R.: On the shape of a set of points in the plane. *IEEE Transactions on Information Theory* 29(4), 551–559 (1983)
4. Edelsbrunner, H.: Weighted alpha shapes. Tech. rep., Champaign, IL, USA (1992)
5. Jarvis, R.A.: Computing the shape hull of points in the plane. In: *Comput. Soc. Conf. Pattern Recognition and Image Processing*, pp. 231–241 (1977)
6. Maillot, Y., Adam, B., Melkemi, M.: Shape reconstruction from unorganized set of points. In: Campilho, A., Kamel, M. (eds.) *ICIAR 2010. LNCS*, vol. 6111, pp. 274–283. Springer, Heidelberg (2010)
7. Mandal, D.P., Murthy, C.A.: Selection of alpha for alpha-hull in R2. *Pattern Recognition* 30(10), 1759–1767 (1997)
8. Melkemi, M.: A-shapes of a finite point set. In: *SCG 1997: Proceedings of the Thirteenth Annual Symposium on Computational Geometry*, pp. 367–369. ACM, New York (1997)
9. Sharvit, D., Chan, J., Tek, H., Kimia, B.B.: Symmetry-based indexing of image databases. *Visual Communication and Image Representation* 9, 366–380 (1998)

Steganalysis of LSB Matching Based on the Statistical Analysis of Empirical Matrix

Hamidreza Dastmalchi and Karim Faez

Amirkabir University of Technology (Tehran Polytechnic)
Tehran, Iran
r.dastmalchi@gmail.com

Abstract. In this paper, the statistical effect of embedding data on Empirical Matrix (EM) of original and differential images is investigated and a novel steganalysis method, targeted at LSB Matching is proposed. It can be mathematically proven, that embedding data in a digital image, causes its empirical matrix and, also the empirical matrixes of its differential images to smooth. Therefore, the high frequency components of an image empirical matrix are omitted due to data hiding which motivates us to extract the radial moments of EM characteristic function as discriminative features for classification. Support Vector Machine with Gaussian kernel is adopted as an appropriate classifier in classification. Experimental results show that the extracted features are highly efficient in attacking LSB Matching.

Keywords: steganalysis, data hiding, empirical (co-occurrence) matrix, characteristic function.

1 Introduction

In the past few years, information hiding has drawn an increasing attention in the field of information security and hidden communication. Information hiding is to hide data in a cover medium in a way that no doubt arises from communication channel observers. Steganography is one of the main typical applications of information hiding which is used for covert communication. The main goal of steganography in digital images is to embed as much information as possible in an ordinary image without causing any noticeable change in neither perceptual nor statistical aspects of the original image. In contrast to steganography, steganalysis is the art of detecting whether a cover medium contains hidden data or not. With the increasing demand for network security, various steganalysis methods have been developed to avoid covert illegal communications. Generally, two types of steganalysis algorithms exist. Target steganalysis algorithms, are designed to attack a special steganography method, while universal algorithms are designed to attack a wide variety of algorithm.

Steganalysis, particularly the universal type can be considered as a pattern recognition problem in which the stego and cover images, represented by some

discriminative features, should discriminate from each other. Although the universal steganalysis algorithms are planned to detect the presence of the message in most of steganography algorithms, their most important goal is to reliably attack LSB Matching as the most prevalent data hiding techniques. In [2], Harmsen et al. proposed a novel steganalysis method called Histogram Characteristic Function Center of Mass (HCFCOM) in which the moments of Histogram Characteristic Function (HCF) were used as features for classification. But this algorithm, suffers from an insufficiency of discriminative features for classification.

In [3], X. Chen et al. proposed an innovative universal steganalysis algorithm based on the features extracted from the co-occurrence matrix (Empirical Matrix) of an image. They claimed that the concentration effect of the empirical matrix is reduced after embedding data. According to this effect of data hiding, they offered to construct the projected histogram from the empirical matrix and extract the multi-order moments of the projected histogram and its characteristic function as classification features.

Afterwards, some other new steganalysis techniques based on empirical matrixes of image have been proposed such as [6,9]. All the Empirical based methods are based on the assumption that embedding data, causes the empirical matrixes of an image to smooth. No mathematical proof has been given to verify smoothing effect of data hiding on empirical matrix, yet. This paper presents a statistical analysis which mathematically verifies the smoothing effect of data hiding on empirical matrixes of digital images. At the rest of paper, an effective steganalysis algorithm is proposed on this mathematical analysis. The Algorithm is blind in attacking different data hiding techniques; however its design is based on the empirical matrix alterations caused by LSB Matching. Therefore, its performance in attacking LSB Matching is investigated and the results are compared with some other efficient steganalysis algorithms.

2 The Statistical Analysis of Empirical Matrix

Universal steganalysis is considered as two-class pattern recognition problem. In other words, some appropriate discriminative features should be extracted from the cover and stego images and a classifier must be trained on a large variety of training images to classify the stego images from the cover ones.

The extracted features must have some specific characteristics which qualify them to be used in steganalysis. First, they must be very sensitive to data embedding. Second, the discriminative features must be relatively independent from the textural contents of images because of the diversity of the images used for embedding data.

Empirical matrix is a very good statistical representation of images which reflects distortions caused by data hiding, efficiently. In this paper, we decide to exploit the empirical matrixes of differential images for extracting appropriate features. The mathematical analysis of data hiding effects on empirical matrixes of original and differential images has been given in sections 2.2 and 2.3. The mathematical analysis verifies the propriety of the exploited features in the proposed steganalysis method.

2.1 The Statistical Definition of Empirical Matrix

The empirical matrix also referred as co-occurrence matrix of an image, reflects the joint distribution of two adjacent pixels. For image I , with L different gray levels, the EM is defined as follow:

$$H_{r,\theta}(i, j) = P(I(x_1, y_1) = i, I(x_2, y_2) = j) \quad (1)$$

$$\begin{cases} x_2 = x_1 + r \cdot \cos(\theta) \\ y_2 = y_1 + r \cdot \sin(\theta) \end{cases}$$

where $P(i, j)$ is a function that computes the number of co-occurring values at a given offset all over the image. r and θ , represent the type of the adjacency for the neighboring pixels. For each r and θ chosen, a different empirical matrix can be obtained and used for feature extraction.

2.2 The Statistical Distortion in Empirical Matrix of Original Image Caused by Data Hiding

In the previous works [3],[6] and [9], it was asserted that embedding data into a digital image, causes its empirical matrix (EM) to smooth. In this section, a mathematical analysis of EM, based on joint probabilities of adjacent pixels, is developed which clearly proves the smoothing effect of Steganography on EM in the case of using LSB Matching as the data hiding technique.

Harmsen and Pearlman [2] showed that data hiding can be modeled as additive noise. The embedding noise probability mass function (PMF) represented by $f_{\Delta}(n)$ is the distribution of the additive noise which is the probability that a pixel alters by n after data hiding. Harmsen, showed that the PMF of embedding noise for LSB Matching technique (represented by $f_{\Delta}(n)$) is a function of embedding rate (α) as follows:

$$f_{\Delta}(n) = \left(\frac{\alpha}{4}\right) \cdot \delta(n+1) + \left(\frac{\alpha}{2}\right) \cdot \delta(n) + \left(\frac{\alpha}{4}\right) \cdot \delta(n-1) \quad (2)$$

By considering the PMF of embedding noise indicated in (2) and by applying probability rules, it can be mathematically proved that embedding data in a digital image with LSB Matching algorithm has a smoothing effect on empirical matrix. To say it more clearly, data hiding effect on EM is equivalent to applying a low-pass filter on the empirical matrix. We call this filter, "Embedding Effect Filer" to emphasize the filtering effect of data hiding.

By dividing the values of empirical matrix elements to the number of the pixels, the joint probability of adjacent pixels is obtained. The joint probabilities of adjacent pixels (x_1 and x_2) of the cover and stego images are represented by $P_c(i, j)$ and $P_s(i, j)$, respectively and are related as follows:

$$P_s(x_{1S} = i, x_{2S} = j) = \sum_{m,n} P_c(x_{1C} = i + m, x_{2C} = j + n) \times f_{\Delta}(-m) \times f_{\Delta}(-n) \quad (3)$$

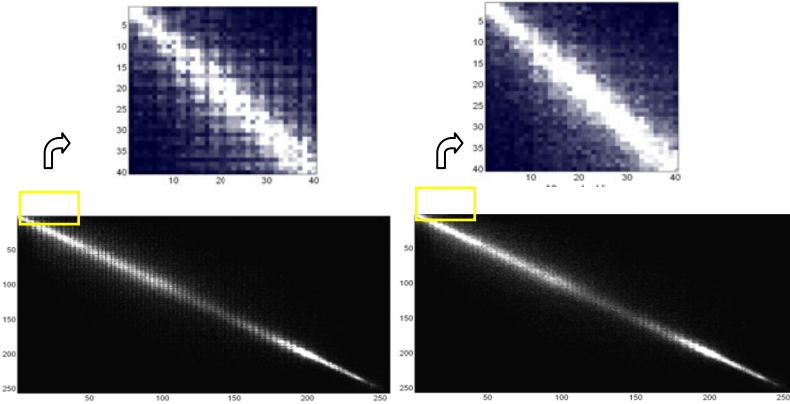


Fig.1. The empirical matrixes of the cover and stego images. The left one is the empirical matrix before embedding data and the right one is the EM after data hiding.

In (3), x_{1C} and x_{2C} are the adjacent pixels prior to data hiding. x_{1S} and x_{2S} are the corresponding adjacent pixels after embedding data. This equation is obtained by considering the fact that the pixels can be changed by maximally one level due to embedding data. Therefore, it is possible that the adjacent pixels in cover (x_{1C} and x_{2C}) differ from their corresponding adjacent pixels in stego (x_{1S} and x_{2S}) by m and n ($|m|, |n| < 1$) and the stego noise (equal to $-m$ and $-n$) compensates for the difference. Since, the embedding noise is independent of the cover image and also because of the independency between the noise sequences, all the probabilities are multiplied as in (3). By multiplying both sides of (3) to M (number of Image pixels) and omitting x indices for more simplicity, equation (4) is obtained:

$$H_s(i, j) = \sum_{n=-1}^{+1} \sum_{m=-1}^{+1} (H_c(i - m, j - n) \cdot f_{\Delta}(m) \cdot f_{\Delta}(n)) \tag{4}$$

in which $H_s(i, j)$ and $H_c(i, j)$ represents for the empirical matrix, after and prior to data hiding, respectively. We can replace right side of (4) by a convolution relationship between $H_c(i, j)$ and a defined “embedding effect filter” represented by $f_{\Delta}(i, j)$ as follows:

$$H_s(i, j) = H_c(i, j) * f_{\Delta}(i, j) \tag{5}$$

In which,

$$f_{\Delta}(i, j) = f_{\Delta}(i) \cdot f_{\Delta}(j)$$

By applying Fourier transform to both sides of (5), the convolution in spatial domain is replaced by multiplication in frequency domain:

$$H_s(u, v) = H_c(u, v) F_{\Delta}(u, v) \tag{6}$$

in which $F_{\Delta}(u)$ represents for the Fourier transform of the embedding noise Probability Mass Function ($f_{\Delta}(n)$).

The Fourier Transform of $f_{\Delta}(n)$ is non-increasing which causes the embedding effect filter $f_{\Delta}(i, j)$ to be a low pass filter as below:

$$F_{\Delta}(u, v) = F_{\Delta}(u) \cdot F_{\Delta}(v) = (\alpha/2 + \alpha/2 \cos(u)) \cdot (\alpha/2 + \alpha/2 \cos(v)) \quad (7)$$

As a result, embedding data in a digital image is equal to applying a low pass filter (with size of 3×3) to the empirical matrix. Figure (1), shows the EM of an image from Corel database before and after embedding data.

As it is clearly shown in figure (1), the EM of the cover image is smoother than the EM of stego due to the smoothing effect of the filter obtained in (7).

2.3 The Statistical Distortion in Empirical Matrix of Differential Images Caused by Data Hiding

Instead of extracting the features from the EM of the original image; we decide to use the EM of differential images for feature extraction. It can be shown that the tiny variations in EM caused by data hiding are magnified by using the differential images. Differential images are constructed in 3 directions: horizontal, vertical and diagonal as in (8):

$$\begin{aligned} I_H(i, j) &= I(i, j) - I(i + 1, j) \\ I_V(i, j) &= I(i, j) - I(i, j + 1) \\ I_D(i, j) &= I(i, j) - I(i + 1, j + 1) \end{aligned} \quad (8)$$

Using the additive noise model proposed by Harmsen [2] and probability rules, we can obtain the PMF of the embedding noise in the differential images as follow:

$$f_{\Delta}^{diff}(n) = \sum_{i=-1}^{+1} f_{\Delta}(i) \cdot f_{\Delta}(n - i) \quad (9)$$

In which, $f_{\Delta}(n)$ and $f_{\Delta}^{diff}(n)$ are the PMF of the embedding noise in the original and differential images, respectively.

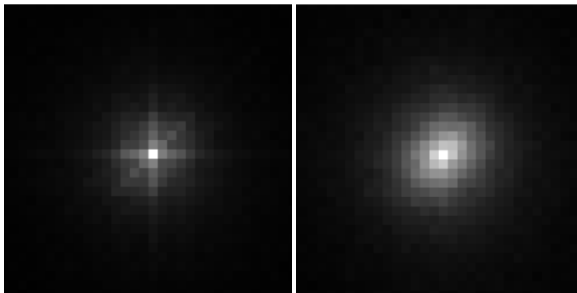


Fig.2. The empirical matrix of horizontal differential image before (left) and after data hiding (right)

According to (9), the PMF of the embedding noise in a differential image is the convolution of embedding noise PMF in the original image with itself. As a result, $f_{\Delta}^{diff}(n)$ is a low-pass filter with a lower cut-off frequency which has a stronger smoothing effect than the one in the original images. Figure 2 shows the empirical matrix of a differential image after and before embedding. This figure depicts the strong smoothing effect of data hiding on differential empirical matrix.

Smoothing effect of data hiding motivates us to use some appropriate measures which reflect the frequency content of differential empirical matrixes as discriminative features for our steganalysis system.

3 Feature Extraction

We decide to extract the features for the steganalysis system from the empirical matrixes of differential images in frequency domain. The magnitude of 2-D DFT of an empirical matrix is called Empirical Matrix Characteristic Function (*EMCF*) and its moments are utilized as the appropriate features.

Generally, two frequency variables exist in a 2-D DFT (u and v) and the moments of *EMCF* can be calculated with respect to each of the frequency variables. In order to compute frequency components along both frequency variables, radial moments are suggested to be used as discriminative features. The n th order radial moment of the *EMCF* is defined as follows:

$$M_{r,\theta}^n = \frac{\sum_{v=-\pi}^{v=\pi} \sum_{u=-\pi}^{u=\pi} |H_{r,\theta}^{diff}(u,v)| \cdot (u^2 + v^2)^{n/2}}{\sum_{v=-\pi}^{v=\pi} \sum_{u=-\pi}^{u=\pi} |H_{r,\theta}^{diff}(u,v)|} \quad (10)$$

where $|H_{r,\theta}^{diff}(u,v)|$ is the *EMCF* of a differential image and the multiplying term $(u^2 + v^2)^{n/2}$, is the norm of two frequency variables to the power of n . In our experiments, we use $H_{r,\theta}$ in 9 different adjacencies (three angles and for each angle three steps) represented by the following pairs of (r, θ) :

$$\{(1,0), (2,0), (3,0), (\sqrt{2}, \pi/4), (2\sqrt{2}, \pi/4), (3\sqrt{2}, \pi/4), (1, \pi/2), (2, \pi/2), (3, \pi/2)\}$$

Therefore, for each test image, we obtain 3 differential images and for each differential image, nine empirical matrixes are computed. Finally, three radial moments of *EMCF* are extracted from each of them. As a result, we have $(81=3*9*3)$ features for classification. We adopt the prediction-error image proposed in [5] to reduce the miscellaneous information and repeat the above feature extraction steps for the prediction-error image, thus there is 81 other features. Therefore, for each image, 162 discriminative features are extracted and used for classification.

4 Experimental Results

We test our proposed method on a 4000-member subset of Corel database. We randomly select 2500 images of the subset for training and the rest 1500 images for

the test. SVM classifier with Gaussian kernel is adopted for classifying the images.

Fig. 3 shows the performance of the algorithm in attacking LSB Matching steganography at different embedding rates compared with three common steganalysis algorithms proposed in [1,3,8].

The detection rate of the steganalysis methods drop with decreasing the embedding rate. The proposed steganalysis algorithm shows a better performance than the other steganalysis algorithms particularly in low embedding rates.

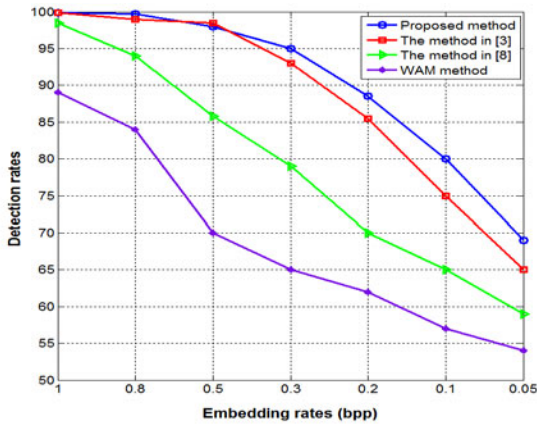


Fig. 3. The recognition rate of different steganalysis techniques in attacking LSB matching with respect to different embedding rates

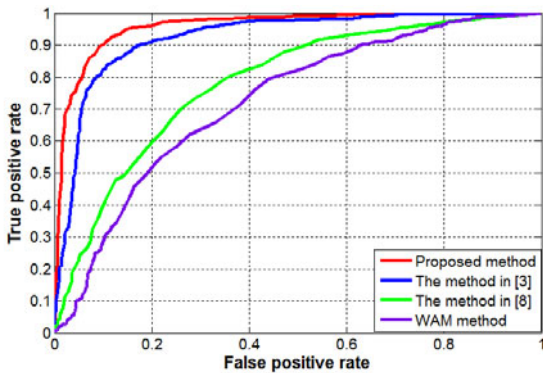


Fig. 4. The ROC curves related to different steganalysis techniques in embedding rate of 0.2 bpp.

According to figure 3, the proposed algorithm shows a significantly better performance than WAM algorithm [1] and the method of Xuan [8] which are both based on the distortions caused by data hiding on the statistical characteristics of wavelet sub-bands. The detection rates of the proposed algorithm is close to the

recognition rate of the method proposed in [3] by Chen in embedding rates higher than 0.3. However, the proposed system outperforms the one in [3] in low embedding rates. In Fig. 4, we give receiver operating characteristic (ROC) curves, showing how false-positive and false-negative errors tradeoff as the detection threshold is varied in the embedding rate of 0.2. According to this figure, the proposed algorithm surpasses the other algorithms as it gives a higher true detection rate in each false positive rate.

5 Conclusion

In this paper, a mathematical analysis of empirical matrix (EM) of original and differential images is proposed. According to the statistical analysis, embedding data into an image by LSB Matching algorithm causes the EM of the original and differential images to smooth. An effective steganalysis method for attacking LSB Matching is also proposed based on the smoothing effect of data hiding. The experimental results show that the proposed method is a promising algorithm in steganalysis field. Our work in the near future is to investigate the same features in wavelet sub-bands.

References

1. Goljan, M., Fridrich, J., Holotyak, T.: New Blind Steganalysis and its Implications. In: Security, Steganography, and Watermarking of Multimedia Contents, pp. 1–13 (2006)
2. Harmsen, J.J., Pearlman, W.A.: Steganalysis of additive noise modelable information hiding. In: Proceedings of the SPIE, Security, Steganography, and Watermarking of Multimedia Contents V, vol. 5020, pp. 131–142 (2003)
3. Chen, X.C., Wang, Y.H., Tan, T.N., Guo, L.: Blind image steganalysis based on statistical analysis of empirical matrix. In: Proceedings of 18th International Conference on Pattern Recognition, vol. 3, pp. 1107–1110 (2006)
4. Farid, H.: Detecting hidden messages using higher-order statistical models. In: Proceedings of IEEE International Conference on Image processing, vol. 2, pp. 905–908 (2002)
5. Yun, Shi, Q., et al.: Image Steganalysis Based on Moments of Characteristic Functions Using Wavelet Decomposition, Prediction-Error Image, and Neural Network. In: ICME, pp. 269–272 (2005)
6. Ziwen, S., Maomao, H., Chao, G.: Steganalysis Based on Co-occurrence Matrix of Differential Image. In: Proceedings of 6th Information Hiding Workshop. LNCS, pp. 1097–1100 (2008)
7. Lie, W.N., Lin, G.-S.: A feature-based classification technique for blind image steganalysis. *IEEE Trans. On Multimedia* 7(6), 1007–1020 (2005)
8. Shi, Y.Q., Xuan, G.R.: Effective steganalysis based on statistical moments of wavelet characteristic function. In: Proceedings of IEEE International Conference on Information Technology: Coding and Computing, pp. 768–773 (2005)
9. Liu, Z., et al.: an Effective Steganalysis Based on Statistical Moments of Differential Characteristic Function. In: Computational Intelligence and Security, pp. 1195–1198 (2006)

Infinite Generalized Gaussian Mixture Modeling and Applications

Tarek Elguebaly and Nizar Bouguila

Concordia Institute for Information Systems Engineering, Concordia University,
Montreal, Canada, Qc, H3G 2W1

t_elgue@encs.concordia.ca, bouguila@ciise.concordia.ca

Abstract. A fully Bayesian approach to analyze infinite multidimensional generalized Gaussian mixture models (IGGM) is developed in this paper. The Bayesian framework is used to avoid model overfitting and the infinite assumption is adopted to avoid the difficult problem of finding the right number of mixture components. The utility of the proposed approach is demonstrated by applying it on texture classification and infrared face recognition, while comparing it to different other approaches.

1 Introduction

Over the last decade, technological advances have brought an explosion of data generation not only in size but also in dimension. These data pose a challenge to standard statistical methods and have received much attention recently. The importance of finding a way to model and analyze multidimensional data lie in their usefulness in wide range of applications such as image processing and computer vision. In recent years a lot of different learning algorithms were developed to recognize complex patterns, and to produce intelligent decisions based on observed data. Mixture models are one of the machine learning techniques receiving considerable attention in different applications. Mixture models are normally used to model complex data sets by assuming that each observation has arisen from one of the different groups or components [1]. In most of the applications, the Gaussian density is used in data analysis. However, many signal processing systems often operate in environments characterized by non-Gaussian and highly peaked sources [2]. Generalized Gaussian distribution (GGD) is considered as a good alternative to the Gaussian due to its shape flexibility which allows it to model a large number of non-Gaussian signals (see, for instance, [3,4,5,2]).

In the recent past, some deterministic approaches have been proposed for the estimation of generalized Gaussian mixture (GGM) models parameters (see, for instance, [6,7,4,8]). Despite the fact that deterministic approaches have dominated mixture models estimation due to their small computational time, many works have demonstrated that these methods have severe problems such as convergence to local maxima, and their tendency to overfit the data [9] especially when data are sparse or noisy. Moreover, another important issue is the difficulty

of getting reliable estimates in case of high dimensional data. With the computational tools evolution, researchers were encouraged to implement and use Bayesian MCMC methods and techniques as an alternative approach. Bayesian methods consider parameters to be random, and to follow different prior distributions (probability distributions). These distributions are used to describe our knowledge before considering the data, as for updating our prior beliefs the likelihood is used. Please refer to [9] for interesting and in depth discussions about the general Bayesian theory. One of the most challenging aspects, when using mixture models, is the estimation of the number of clusters that best describes the data without over or under fitting it. For this purpose, many approaches have been suggested, which can be classified from computational point of view into two groups: deterministic, and Bayesian methods. In this paper we use a Bayesian non-parametric approach based on allowing the number of components to increase to infinity as new data arrive. We describe a Bayesian algorithm for learning IGGM, and demonstrate its effectiveness by applying it to two real applications namely image texture classification and infrared face recognition.

The remainder of this paper is organized as follows. The next section describes our Bayesian learning approach. Section 3 presents the complete algorithm used for learning the model parameters. In section 4, we assess the performance of our model on different applications while comparing it to other models. Our last section is devoted to the conclusion.

2 Learning of the IGGM Model

2.1 The Mixture Model

If a d -dimensional $\mathbf{X} = (X_1, \dots, X_d)$ follows a GGD, then:

$$P(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\alpha}, \boldsymbol{\beta}) = \prod_{k=1}^d \frac{\beta_k \alpha_k}{2\Gamma(1/\beta_k)} e^{-(\alpha_k |X_k - \mu_k|^{\beta_k})} \quad (1)$$

where $\boldsymbol{\mu} = (\mu_1, \dots, \mu_d)$, $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d)$, and $\boldsymbol{\beta} = (\beta_1, \dots, \beta_d)$ are the mean, the inverse scale, and the shape parameters. Let $\mathcal{X} = (\mathbf{X}_1, \dots, \mathbf{X}_N)$ be a set of N iid vectors assumed to arise from a GGM with M components:

$$P(\mathcal{X}|\Theta) = \sum_{j=1}^M P(\mathbf{X}|\boldsymbol{\mu}_j, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j) p_j \quad (2)$$

where $\{p_j\}$ are the mixing proportions which must be positive and sum to one. The set of parameters of the mixture with M components is defined by $\Theta = (\{\boldsymbol{\mu}_j\}, \{\boldsymbol{\alpha}_j\}, \{\boldsymbol{\beta}_j\}, \{p_j\})$. We introduce stochastic indicator variables, Z_i , one for each observation, whose role is to encode to which component the observation belongs. In other words, Z_{ij} , the unobserved or missing vector, equals 1 if \mathbf{X}_i belongs to class j and 0, otherwise. The complete-data likelihood for this case is then:

$$P(\mathcal{X}, Z|\Theta) = \prod_{i=1}^N \prod_{j=1}^M (P(\mathbf{X}_i|\boldsymbol{\xi}_j) p_j)^{Z_{ij}} \quad (3)$$

where $Z = \{Z_1, Z_2, \dots, Z_N\}$, and $\xi_j = (\boldsymbol{\mu}_j, \boldsymbol{\alpha}_j, \boldsymbol{\beta}_j)$. Bayesian MCMC simulation methods are based on the well-known Bayesian formulae:

$$\pi(\Theta|\mathcal{X}, Z) = \frac{\pi(\Theta)P(\mathcal{X}, Z|\Theta)}{\int \pi(\Theta)P(\mathcal{X}, Z|\Theta)} \propto \pi(\Theta)P(\mathcal{X}, Z|\Theta) \tag{4}$$

where (\mathcal{X}, Z) , $\pi(\Theta)$ and $\pi(\Theta|\mathcal{X}, Z)$ are the complete data, the prior information about the parameters and the posterior distribution, respectively. Having $\pi(\Theta|\mathcal{X}, Z)$ we can simulate our model parameters Θ , rather than computing them.

For the $\{p_j\}$, we know that $(0 \leq p_j \leq 1$ and $\sum_{j=1}^M p_j = 1)$, then the typical choice, as a prior, is a symmetric Dirichlet distribution with parameter η/M . As for $\pi(Z|p)$ we have:

$$\pi(Z|p) = \prod_{j=1}^M \pi(Z_i|p) = \prod_{i=1}^N \prod_{j=1}^M p_j^{Z_{ij}} = \prod_{j=1}^M p_j^{n_j} \tag{5}$$

where $n_j = \sum_{i=1}^N \mathbf{I}_{Z_{ij}=1}$. Then using the standard Dirichlet integral, we may integrate out the mixing proportions and write the prior directly in terms of the indicators:

$$\pi(Z|\eta) = \frac{\Gamma(\eta)}{\Gamma(\eta + N)} \prod_{j=1}^M \frac{\Gamma(\eta/M + n_j)}{\Gamma(\eta/M)} \tag{6}$$

In order to be able to use Gibbs sampling for the missing vector, Z , we need the conditional prior for a single indicator given all the others; this can be easily obtained from Eq. 6 by keeping all but a single indicator fixed:

$$\pi(Z_i = j|\eta, Z_{-i}) = \frac{n_{-ij} + \eta/M}{N - 1 + \eta} \tag{7}$$

where the subscript $-i$ indicates all indexes except i . Note that n_{-ij} is the number of observations, excluding \mathbf{X}_i , in cluster j . For the parameters ξ , we assign independent Normal prior with δ, ε^2 as the mean and variance for the distributions means $(\boldsymbol{\mu}_j)$, respectively. Independent Gamma prior with ι, ρ as the shape and rate parameters, respectively, is assigned for the inverse scale $\boldsymbol{\alpha}_j$. For the shape parameter, $\boldsymbol{\beta}_j$, we used independent Gamma prior with κ, ς as the shape and rate parameters, respectively [10]. Thus, the posterior distributions for $\boldsymbol{\mu}_j, \boldsymbol{\alpha}_j$, and $\boldsymbol{\beta}_j$ are given by:

$$P(\boldsymbol{\mu}_j|Z, \mathcal{X}) \propto \prod_{k=1}^d \frac{1}{\varepsilon} e^{-\frac{(\mu_{jk} - \delta)^2}{2\varepsilon^2}} \times \prod_{k=1}^d e^{\sum_{Z_{ij}=1} (-\alpha_{jk} |X_{ik} - \mu_{jk}|)^{\beta_{jk}}} \tag{8}$$

$$P(\boldsymbol{\alpha}_j|Z, \mathcal{X}) \propto \prod_{k=1}^d \frac{\alpha_{jk}^{\iota-1} \rho^\iota e^{-\rho\alpha_{jk}}}{\Gamma(\iota)} \times \prod_{k=1}^d \left[\alpha_{jk} \right]^{n_j} e^{\sum_{Z_{ij}=1} (-\alpha_{jk} |X_{ik} - \mu_{jk}|)^{\beta_{jk}}} \tag{9}$$

$$P(\beta_j|Z, \mathcal{X}) \propto \prod_{k=1}^d \frac{\beta_{jk}^{\kappa-1} \zeta^\kappa e^{-\zeta \beta_{jk}}}{\Gamma(\kappa)} \times \prod_{k=1}^d \left[\frac{\beta_{jk}}{\Gamma(1/\beta_{jk})} \right]^{n_j} e^{\sum_{i,j=1} (-\alpha_{jk} |X_{ik} - \mu_{jk}|)^{\beta_{jk}}} \tag{10}$$

In order to have a more flexible model, we introduce an additional hierarchical level by allowing the hyperparameters to follow some selected distributions. The hyperparameters, δ and ε^2 associated with the μ_j are given Normal and Inverse Gamma priors with parameters (ϵ, χ^2) and (φ, ϱ) , respectively. Thus,

$$P(\delta|\dots) \propto P(\delta|\epsilon, \chi^2) \prod_{j=1}^M P(\mu_j|\delta, \varepsilon^2) \propto e^{\frac{-(\delta-\epsilon)^2}{2\chi^2}} \times \prod_{j=1}^M \prod_{k=1}^d e^{\frac{-(\mu_{jk}-\delta)^2}{2\varepsilon^2}} \tag{11}$$

$$P(\varepsilon^2|\dots) \propto P(\varepsilon^2|\varphi, \varrho) \prod_{j=1}^M P(\mu_j|\delta, \varepsilon^2) \propto \frac{\exp(-\varrho\varepsilon^2)}{\varepsilon^{2(\varphi+1)}} \left[\frac{1}{\varepsilon} \right]^{Md} \times \prod_{j=1}^M \prod_{k=1}^d e^{\frac{-(\mu_{jk}-\delta)^2}{2\varepsilon^2}} \tag{12}$$

The hyperparameters ι and ρ associated with the α_j are given inverse Gamma and Gamma priors with parameters (ϑ, ϖ) and (τ, ω) , respectively. Thus,

$$P(\iota|\dots) \propto P(\alpha_\alpha|\vartheta, \varpi) \prod_{j=1}^M P(\alpha_j|\iota, \rho) \propto \frac{\exp(-\varpi/\iota)}{\iota^{\vartheta+1}} \left[\frac{\rho^\iota}{\Gamma(\iota)} \right]^{Md} \times \prod_{j=1}^M \prod_{k=1}^d \alpha_{jk}^{\iota-1} e^{-\rho\alpha_{jk}} \tag{13}$$

$$P(\rho|\dots) \propto P(\beta_\alpha|\tau, \omega) \prod_{j=1}^M P(\alpha_j|\iota, \rho) \propto \rho^{\tau-1} e^{-\omega\rho} \left[\rho^\iota \right]^{Md} \times \prod_{j=1}^M \prod_{k=1}^d \alpha_{jk}^{\iota-1} e^{-\rho\alpha_{jk}} \tag{14}$$

The hyperparameters κ and ς associated with the β_j are given inverse Gamma and Gamma priors with parameters (λ, ϕ) and (ν, ψ) , respectively. Thus,

$$P(\kappa|\dots) \propto P(\kappa|\lambda, \phi) \prod_{j=1}^M P(\beta_j|\kappa, \varsigma) \propto \frac{\exp(-\phi/\kappa)}{\kappa^{\lambda+1}} \left[\frac{\varsigma^\kappa}{\Gamma(\kappa)} \right]^{Md} \times \prod_{j=1}^M \prod_{k=1}^d \beta_{jk}^{\kappa-1} e^{-\varsigma\beta_{jk}} \tag{15}$$

$$P(\varsigma|\dots) \propto P(\varsigma|\nu, \psi) \prod_{j=1}^M P(\beta_j|\kappa, \varsigma) \propto \varsigma^{\nu-1} e^{-\psi\varsigma} \left[\varsigma^\kappa \right]^{Md} \times \prod_{j=1}^M \prod_{k=1}^d \beta_{jk}^{\kappa-1} e^{-\varsigma\beta_{jk}} \tag{16}$$

2.2 The IGGM Model

So far, we have considered M to be a fixed quantity. In this section, we overcome this obstacle by assuming that $M \rightarrow \infty$ in Eq. 7 which gives us

$$\pi(Z_i = j|\eta, Z_{-i}) = \begin{cases} \frac{n_{-ij}}{N-1+\eta}; & \text{if } n_{-ij} > 0 \text{ (cluster } j \in \mathcal{R}) \\ \frac{\eta}{N-1+\eta}; & \text{if } n_{-ij} = 0 \text{ (cluster } j \in \mathcal{U}) \end{cases} \tag{17}$$

where \mathcal{R} and \mathcal{U} are the sets of represented and unrepresented clusters, respectively. Thus, the conditional posterior is obtained by combining this prior with the likelihood of the data:

$$\pi(Z_i = j | \eta, \mu_j, \alpha_j, \beta_j, Z_{-i}, \mathbf{X}_i) = \begin{cases} \frac{n_{-ij}}{N-1+\eta} p(\mathbf{X}_i | \mu_j, \alpha_j, \beta_j); & \text{if } j \in \mathcal{R} \\ \int \frac{\eta}{N-1+\eta} p(\mathbf{X}_i | \mu_j, \alpha_j, \beta_j) p(\mu_j | \delta, \varepsilon^2) p(\alpha_j | \iota, \rho) p(\beta_j | \kappa, \varsigma) d\mu_j d\alpha_j d\beta_j & \text{if } j \in \mathcal{U} \end{cases} \quad (18)$$

The choice of the concentration parameter η is of high importance as it controls the generation frequency of new clusters. We decided to use an Inverse Gamma prior for η :

$$P(\eta | v, \gamma) \sim \frac{\gamma^v \exp(-\gamma/\eta)}{\Gamma(v)\eta^{v+1}} \quad (19)$$

Using the above equation with Eq. 17 we reach the following posterior:

$$P(\eta | \dots) \propto \frac{\gamma^v \exp(-\gamma/\eta)}{\Gamma(v)\eta^{v+1}} \eta^M \prod_{j=1}^N \frac{1}{i-1+\eta} \propto \frac{\gamma^v \exp(-\gamma/\eta)}{\Gamma(v)\eta^{v+1}} \frac{\eta^M \Gamma(\eta)}{\Gamma(N+\eta)} \quad (20)$$

Our hierarchical model can be displayed as a directed acyclic graph (DAG) as shown in Fig. 1.

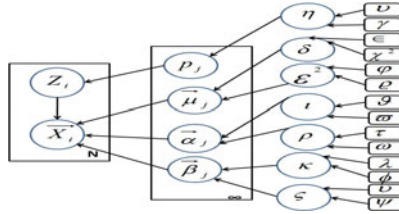


Fig. 1. Graphical Model representation of the Bayesian hierarchical IGGM model. Nodes in this graph represent random variables, rounded boxes are fixed hyperparameters, boxes indicate repetition (with the number of repetitions in the lower right) and arcs describe conditional dependencies between variables.

3 The Complete Algorithm

Having all the conditional posteriors, we can employ a Gibbs sampler with the following steps:

1. Generate Z_i from Eq. 18 then update n_j .
2. Update the number of represented components M .
3. Update the mixing parameters for the represented components by $p_j = \frac{n_j}{N+\eta}$ for $j = 1, \dots, M$, and for the unrepresented components by $p_U = \frac{\eta}{N+\eta}$.
4. Generate the mixture parameters μ_j , α_j , and β_j from Eqs. 8, 9 and 10.
5. Update the hyperparameters δ , ε^2 , ι , ρ , κ , ς , and η from Eqs. 11, 12, 13, 14, 15, 16, 20, respectively.

Note that, for the initialization step we started by assuming that all the vectors are in the same cluster, and we generated the parameters by sampling from their prior distributions. It is quite easy to notice that we cannot simulate directly from these posterior distributions because they are not in well known forms. To solve this problem we applied the well known Metropolis-Hastings (M-H) algorithm given in [11].

4 Experimental Results

In the following applications, we use 5000 iterations for our Metropolis-within-Gibbs sampler (we discarded the first 800 iterations as “burn-in” and kept the rest), and our specific choices for the hyperparameters are

$$(v, \gamma, \epsilon, \chi^2, \varphi, \varrho, \vartheta, \varpi, \tau, \omega, \lambda, \phi, \nu, \psi) = (2, 0.2, 1, 0.5, 2, 5, 2, 5, 2, 0.2, 2, 5, 2, 0.2)$$

4.1 Categorization of Texture Images

In this application we are interested by the categorization of texture images which is important in the case of content-based image retrieval, for instance. In order to determine the vector of characteristics for a given texture, we use set of features derived from the image correlogram [12], by considering four neighborhoods and directions: $(1; 0)$, $(1, \pi/4)$, $(1, \pi/2)$, and $(1, 3\pi/4)$, from which we derive eight features: mean, variance, energy, correlation, entropy, contrast, homogeneity, and cluster prominence [13]. Thus, each image is represented by a 32-dimensional vector. Finally, we apply two methods, our IGGM and the infinite Gaussian mixture (IGM) [14], in order to categorize the images.

We perform our experiments using the Vistex texture data set [4]. Six homogeneous texture groups (Bark, Fabric, Food, Metal, Water, and Sand) are considered. We use four 512×512 images from each of the Bark, Fabric, and Metal texture groups, and six 512×512 from each of the Food, Water, and Sand texture groups, then we divide each image into sixty four 64×64 subimages. Now, we have a total of 1,920 sub-images: 256 sub-images for each class in the first three groups, and 384 sub-images for each class in the second three groups. Examples of images from each of the six categories are shown in figure 2.

The IGGM mixture favored 6 categories which is the case here. The IGM, however, classified the texture images into 7 clusters where the 7th component had a very small probability of 0.0276. In order to be able to compare both methods, we supposed that we obtained the right number of clusters in the case of the IGM. The confusion matrices for both methods are given in tables 1.a and 1.b. As shown the total number of misclassified images in the case of IGGM is 36 which identifies a high accuracy of 98.12%. The accuracy of the IGM was 93.80%, as it misclassified 119 images.

¹ MIT Vision and Modeling Group (<http://vismod.www.media.mit.edu>)

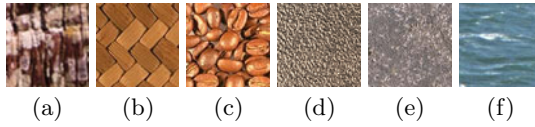


Fig. 2. Sample images from each group. (a) Bark, (b) Fabric, (c) Food, (d) Metal, (e) Sand, (f) Water.

Table 1. Confusion matrix for texture categorization using (a) IGGM and (b) IGM

	Bark	Fabric	Metal	Food	Sand	Water		Bark	Fabric	Metal	Food	Sand	Water
Bark	255	0	0	0	1	0	Bark	241	0	0	1	6	5
Fabric	0	248	0	8	0	0	Fabric	2	238	0	6	2	2
Metal	0	0	252	0	0	4	Metal	0	2	237	3	0	4
Food	0	6	0	378	0	0	Food	0	7	3	362	0	2
Sand	3	0	0	0	380	1	Sand	5	0	2	0	363	3
Water	3	2	2	4	2	371	Water	4	1	2	3	1	360

(a)

(b)

4.2 Infrared Face Recognition

Recently, different studies have shown that thermal IR offers a promising alternative to visible imagery for handling variations in face appearance [15]. Figure 3 shows visual and thermal image characteristics of faces with variations in illumination and facial expression. Although illumination and facial expression significantly change the visual appearance of the face, thermal characteristics of the face remain nearly invariant. Several approaches have been proposed to analyze and recognize infrared faces and can be divided into two main groups: appearance-based and feature-based methods. While appearance-based methods focus on the global properties of the face, feature-based methods explore the facial features (ex. eyes, mouth) statistical and geometrical properties [16]. Many of these approaches, however, suppose that the extracted infrared face features are Gaussian which is not generally an appropriate assumption. We propose then, in this section, an appearance-based approach using IGGM. We are



Fig. 3. Visual and thermal image characteristics of faces with variations in illumination

considering face recognition as an image classification problem by trying to classify to which person this image belongs. For feature extraction step we have employed both the edge orientation histograms [17] and the co-occurrence matrices which capture the local spatial relationships between gray levels [18]. Figure 4 shows 3 face images, where the first two images are for the same person taken from different poses, and the third image is for another person. It is quite clear that the first two images have very close edge-orientation histograms compared to the third image. In our experiments, we have considered the following four co-occurrence matrices: $(1; 0)$, $(1, \pi/4)$, $(1, \pi/2)$, and $(1, 3\pi/4)$, respectively [19]. For each co-occurrence matrix we derived the following features: mean, variance, energy, correlation, entropy, contrast, homogeneity, and cluster prominence [19]. Besides, the edge directions are quantized into 72 bins of 5° each. Using the co-occurrence matrices and the histogram of edge directions each image was represented by a 104-dimensional vector.

In our experiments, we performed face recognition using images from the Iris thermal face database which is a subset of the Object Tracking and Classification Beyond the Visible Spectrum (OTCBVS) database. First we used 1320 images of fifteen persons not wearing glasses. Knowing that in IR imaging thermal radiation cannot transmit through glasses because glasses severely attenuate electromagnetic wave radiation beyond 2.4 mm, we decided to investigate if our algorithm will be capable to identify persons with glasses, so we added 880 images of eight persons with glasses. For both experiments we used 11 images

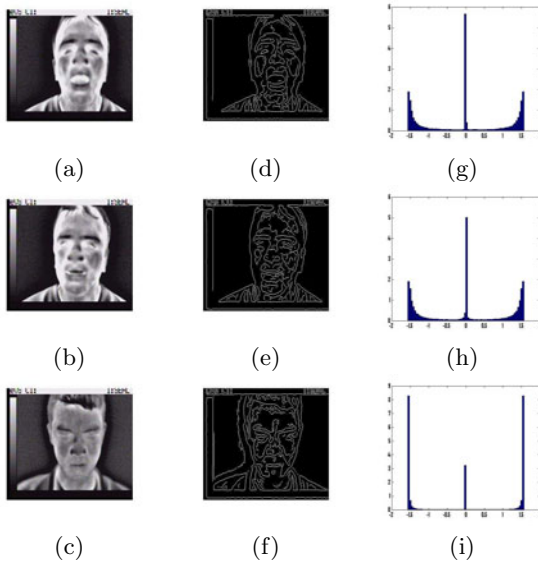


Fig. 4. 3 different images of two different persons with their corresponding shape images and corresponding shape histograms, (a)-(c) show three database images, (d)-(f) show the corresponding edge images, (g)-(i) show the corresponding shape histograms.

Table 2. Accuracies for The seven different methods

	IGGM	EMGGM	IGM	EMGM	PCA	HICA	LICA
Data 1	97.02%	94.20%	86.67%	85.89%	95.58%	95.32%	94.46%
Data 2	96.33%	92.40%	82.54%	82.18%	94.35%	93.99%	92.81%

for each person as training set and the rest as testing set. This gave us 165 and 1155 images for training and testing, respectively, in the first data. The second data set was composed of 253 and 1947 images for training and testing, respectively. In order to validate our algorithm (IGGM) we have compared it with the expectation maximization (EM) one (EMGGM). We also compared it to six other methods namely principal component analysis (PCA) with cosine distance, localized independent component analysis (LICA) with cosine distance, holistic ICA (HICA) with cosine distance as implemented by FastICA [20], IGM and Gaussian mixture models learned with EM (EMGM). Table 2 shows the accuracies for the seven different methods. According to this table it is clear that the IGMM outperforms all other methods which can be explained by its ability to incorporate prior information during classes learning and modeling.

5 Conclusion

We have described and illustrated a Bayesian nonparametric approach based on infinite generalized Gaussian mixtures. We proposed an MCMC algorithm to learn the parameters of this mixture. The effectiveness of the proposed approach has been shown using two important applications namely texture images classification and Infrared face recognition.

Acknowledgment

The completion of this research was made possible thanks to the Natural Sciences and Engineering Research Council of Canada (NSERC).

References

1. McLachlan, G.J., Peel, D.: Finite Mixture Models. Wiley, New York (2000)
2. Elguebaly, T., Bouguila, N.: Bayesian Learning of Finite Generalized Gaussian Mixture Models on Images. *Signal Processing* 91(4), 801–820 (2011)
3. Meignen, S., Meignen, H.: On the Modeling of Small Sample Distributions with Generalized Gaussian Density in a Maximum Likelihood Framework. *IEEE Transactions on Image Processing* 15(6), 1647–1652 (2006)
4. Allili, M.S., Bouguila, N., Ziou, D.: Finite General Gaussian Mixture Modeling and Application to Image and Video Foreground Segmentation. *Journal of Electronic Imaging* 17(1), 1–13 (2008)

5. Elguebaly, T., Bouguila, N.: Bayesian learning of generalized gaussian mixture models on biomedical images. In: Schwenker, F., El Gayar, N. (eds.) ANNPR 2010. LNCS, vol. 5998, pp. 207–218. Springer, Heidelberg (2010)
6. Allili, M.S., Bouguila, N., Ziou, D.: Finite generalized gaussian mixture modeling and applications to image and video foreground segmentation. In: Proc. of the Canadian Conference on Robot and Vision (CRV), pp. 183–190 (2007)
7. Allili, M.S., Bouguila, N., Ziou, D.: A robust video foreground segmentation by using generalized gaussian mixture modeling. In: Proc. of the Canadian Conference on Robot and Vision (CRV), pp. 503–509 (2007)
8. Fan, S.-K.S., Lin, Y.: A Fast Estimation Method for the Generalized Gaussian Mixture Distribution on Complex Images. *Computer Vision and Image Understanding* 113(7), 839–853 (2009)
9. Robert, C.P.: *The Bayesian Choice From Decision-Theoretic Foundations to Computational Implementation*, 2nd edn. Springer, Heidelberg (2007)
10. Robert, C.P., Casella, G.: *Monte Carlo Statistical Methods*, 2nd edn. Springer, Heidelberg (2004)
11. Lewis, S.M., Raftery, A.E.: Estimating Bayes Factors via Posterior Simulation with the Laplace-Metropolis Estimator. *Journal of the American Statistical Association* 90, 648–655 (1997)
12. Huang, J., Kumar, S.R., Mitra, M., Zhu, W.-J., Zabih, R.: Image Indexing Using Color Correlograms. In: Proc. of the IEEE Conference Computer Vision and Pattern Recognition, p. 762 (1997)
13. Randen, T., Husoy, J.H.: Sum and Difference Histograms for Texture Classification. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21(4), 291–310 (1999)
14. Rasmussen, C.E.: The Infinite Gaussian Mixture Model. In: *Advances in Neural Information Processing Systems (NIPS)*, pp. 554–560 (2000)
15. Han, X., Koelling, K.W., Tomasko, D.L., Lee, L.J.: A comparative analysis of face recognition performance with visible and thermal infrared imagery. In: Proc. of the International Conference on Pattern Recognition (ICPR), pp. 217–222 (2002)
16. Arandjelovic, O., Hammoud, R., Cipolla, R.: Multi-sensory face biometric fusion (for personal identification). In: Proc. of the IEEE Workshop on Computer Vision Beyond the Visible Spectrum: Methods and Applications, CVBVS (2006)
17. Jain, A.K., Vailaya, A.: Image retrieval using color and shape. *Pattern Recognition* 29, 1233–1244 (1996)
18. Shanmugam, K., Haralickand, R.M., Dinstein, I.: Texture features for image classification. *IEEE Transactions on Systems, Man, and Cybernetics*, 610–621 (1973)
19. Unser, M.: Filtering for Texture Classification: A Comparative Study. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 8(1), 118–125 (1986)
20. Hyvrinen, A.: The fixed-point algorithm and maximum likelihood estimation for independent component analysis. *Neural Processing Letters* 10, 1–5 (1999)

Fusion of Elevation Data into Satellite Image Classification Using Refined Production Rules

Bilal Al Momani¹, Philip Morrow², and Sally McClean²

¹ Cisco Systems, Galway, Ireland

² School of Computing and Information Engineering

Faculty of Engineering, University of Ulster, Northern Ireland

balmoman@cisco.com, {si.mcclean, pj.morrow}@ulster.ac.uk

Abstract. The image classification process is based on the assumption that pixels which have similar spatial distribution patterns, or statistical characteristics, belong to the same spectral class. In a previous study we have shown how we can improve the accuracy of classification of remotely sensed imagery data by incorporating contextual elevation knowledge in a form of a digital elevation model with the output of the classification process using Dempster-Shafer Theory of Evidence. A knowledge based approach is created for this purpose using suitable production rules derived from the elevation distributions and range of values for the elevation data attached to a particular satellite image. Production rules are the major part of knowledge representation and have the basic form: IF condition THEN Inference. Although the basic form of production rules has shown accuracy improvement, in general, in some cases accuracy can degrade. In this paper we propose a “refined” approach that takes into account the actual “distribution” of elevation values for each class rather than simply the “range” of values to solve the accuracy degradation. This approach is performed by refining the basic production rules used in the previous study taking into account the number of pixels at each elevation within the elevation distribution for each class.

Keywords: Remote sensing, classification, evidence theory.

1 Introduction and Background

Remote sensing is the process of observation and measurement of the earth’s surface from a physical distance. The major component of this process is the interpretation and identification of information and is termed image classification. The traditional approaches for remotely sensed data classification, i.e. “supervised” and “unsupervised” have focused on using spectral data i.e. “within image” data only to perform the classification. However, spectral information has proven to be insufficient for accurate classification in many cases. In addition, there are many other types of external data (e.g. elevation, OS map, and soil type) that are attached to the area of interest, which can be utilized to aid the classification process. In previous studies ([1]), we have shown that contextual data, i.e. “elevation” can be fused with the output of traditional classification algorithms using Dempster-Shafer theory of evidence. A knowledge base of production rules can be generated for this purpose

from the elevation distributions of selected training areas/clusters. This approach has shown significant improvement in classification accuracy. However, in some cases this accuracy can get worse and therefore in this paper we propose a new approach based on modified production rules to address the problem of accuracy degradation. Pixel-based Semi-supervised Classification (SSC) using the “Expectation Maximization” (EM) algorithm [2] via a Gaussian mixture model is performed initially to classify the image. When limited “labelled” data is available then a semi-supervised approach is useful since it can deal with labelled and unlabeled data in a single framework. Moreover, since the correlation between image band data is high then it is appropriate to use a Gaussian model to handle this data [3]. Semi-supervised classification (SSC) has been used in many different applications including text categorization [4], Biological data clustering [5] and for remotely sensed data [3].

1.1 Evidential Theory

Dempster-Shafer theory is a mathematical concept based on belief and plausibility, which is used to combine separate sources of evidence to “calculate the probability of an event” [6]. This theory has been used previously for different applications, e.g. for knowledge discovery [7], for combining different classifiers, for the combination of heterogeneously classified data [8] and for multi-scale data fusion [9]. Data, in this theory, is represented in the form of a mass function, which measures our degree of belief in various propositions or sets of values. The mass function assigns belief to sets, which together form the frame of discernment Ω . The mass function m is defined on subsets of Ω (propositions) as: $m(\phi) = 0$, i.e. the mass function of the null proposition ϕ is always zero; $\sum_{A \subseteq \Omega} m(A) = 1$, i.e. the sum of the masses of all the propositions (A) in the

frame of discernment is one. These definitions indicate that the propositions may be overlapping and therefore provide a lower and upper bound (belief and plausibility) for the probability assigned to a particular proposition. The belief in a proposition is the sum of masses of all propositions contained in it; the plausibility of a proposition is the sum of the masses of all propositions in which it is wholly or partly contained. The belief and plausibility functions are therefore defined by:

$$Bel(A) = \sum_{X \subseteq A} m(X) \quad \text{and} \quad Pls(A) = \sum_{X: X \cap A \neq \emptyset} m(X) \tag{1}$$

The belief and plausibility functions may thus be used to determine the amount of support for a proposition. They may then be used to induce rules based on the mass allocations for various propositions and may be regarded as providing pessimistic and optimistic measures of how strong a rule might be [10]. For example, if we want to classify an area into classes such as grass, soil, etc, we might define the mass function (m) as follows:

Example: $m(\{grass\}) = 0.9$; $m(\{soil\}) = 0.05$; $m(\{grass, soil\}) = 0.05$. Here, we are 90% sure that the area refers to grass, 5% sure that it is soil and 5% sure that it might be either. Hence $Bel(\{grass\}) = 0.9$; $Pls(\{grass\}) = 0.95$. Dempster’s law of combination allows us to combine evidence, in the form of mass functions, from different sources. Let m_1 and m_2 be two mass functions on the frame of discernment

Ω . Then, for any subset $H \subseteq \Omega$, the *orthogonal sum* \oplus of two mass functions of propositions X and Y on H is defined as:

$$(m_1 \oplus m_2)(H) = \frac{\sum_{X \cap Y = H} m_1(X) * m_2(Y)}{1 - \sum_{X \cap Y = \emptyset} m_1(X) * m_2(Y)} \tag{2}$$

The orthogonal sum thus allows two mass functions to be combined into a third mass function, which pools pieces of evidence to support propositions of interest.

1.2 Semi-Supervised Classification

Supervised classification requires the user to have a priori knowledge about the area while unsupervised classification assumes no a priori knowledge about it and therefore the classification process starts “blind”. Overall, it is difficult to obtain adequate and accurate labelled data from satellite images. Therefore, the so called “semi-supervised” classification (SSC) approach which utilises both labelled and unlabelled data has been proposed as a possible solution that copes with these data in a single framework [3]. The basic steps to perform semi-supervised classification are shown in Fig 1.

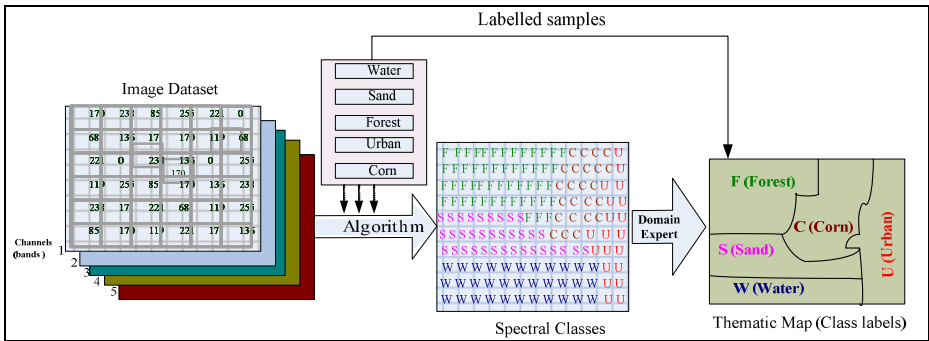


Fig. 1. Basic steps in semi-supervised classification (adapted from [11])

It can be seen from this figure that the classification process makes use of labelled and unlabelled data to perform the classification. The unlabelled data is classified using an appropriate classification algorithm (e.g. Expectation Maximization algorithm) using unsupervised classification. However, the labelled data is excluded from this classification since it already has class labels. The Expectation Maximization (EM) algorithm is used to accomplish semi-supervised classification and is performed in two steps: the E-step is used to locate the posterior probability (conditional probability) of each pixel. The M-step is used to calculate new parameters (mean (μ), covariance matrix (Σ) and proportion (τ)) based on the new posterior probability calculated in the E-step. The two steps keep iterating until convergence is satisfied. A simplified version of the EM algorithm is illustrated in Fig. 2. (explained in more details in [2]). It can be seen from Fig. 2 that we initially define a conditional probability for each pixel for each cluster. For example we might

initialize the conditional probability Z_{ik} for each pixel i of cluster k so that $\sum Z_{ik} = 1$. The M-step then starts for all data, i.e. pixels (labelled and unlabelled). Throughout the E-step the labelled data are excluded from the calculation since they already have class labels.

Initialise “random” *posterior* probabilities Z_{ik} for each pixel i of cluster k so that:

$$\sum Z_{ik} = 1$$

Loop
M-Step: Initialise Cluster Parameters

$$n_k = \sum_{i=1}^n Z_{ik}, \quad \tau_k = \frac{n_k}{n}, \quad \text{where } n_k \text{ is total no. of pixels in}$$

cluster k , n is total no. of pixels and τ is cluster proportion.

$$\mu_k = \frac{Z_{ik} y_i}{n_k}, \quad \Sigma_k = \frac{1}{n_k} \sum_{i=1}^n Z_{ik} (y_i - \mu_k)(y_i - \mu_k)', \quad \text{where } y_i \text{ is the vector value of}$$

pixel i .

E-Step: update the *posterior* probability Z_{ik} as follows:

Labelled data: $Z_{ik} = \begin{cases} 1 & \text{if pixel } i \text{ belongs to cluster } k \\ 0 & \text{otherwise} \end{cases}$

Unlabelled data: (Gaussian mixture model)

$$Z_{ik} = \frac{\tau_k \cdot \frac{1}{\sqrt{2\pi|\Sigma_k|}} \cdot \exp\left\{-\frac{1}{2}(y_i - \mu_k)\Sigma_k^{-1}(y_i - \mu_k)'\right\}}{\sum_{j=1}^k \tau_j \cdot \frac{1}{\sqrt{2\pi|\Sigma_j|}} \cdot \exp\left\{-\frac{1}{2}(y_i - \mu_j)\Sigma_j^{-1}(y_i - \mu_j)'\right\}}$$

Until convergence is satisfied

Fig. 2. Semi-supervised classification using the EM algorithm via Gaussian mixture model

2 Fusing Elevation Data

2.1 Generating Basic Rules

In our previous work, Dempster’s law of combination (equation 2) was used to combine the output of traditional classification algorithms with contextual data in a form of elevation knowledge [1]. This approach is based on using “basic rules”, where all rules are generated based on a fixed confidence (0.95). To illustrate the approach we consider a dataset for part of Nome, Alaska State, USA obtained from the USGS website [12]. The different types of data are: The Landsat Thematic Mapper (TM) (three bands), Digital Elevation Model and Landcover data (i.e. “groundtruth”) as shown in (Fig.3 a, b & c).

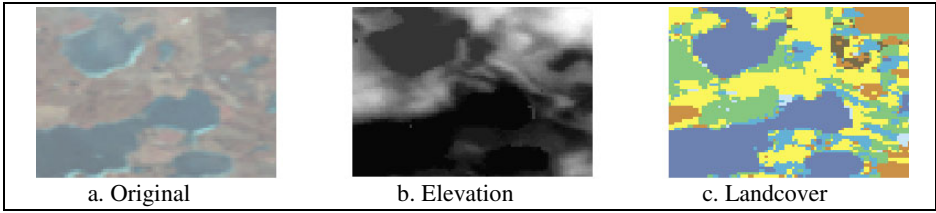


Fig. 3. Study area (a) Original (b) Elevation and (c) Landcover image

The original image (Fig. 3 a) was initially classified using the semi-supervised classification discussed above. We use the landcover image (Fig. 3 c) only for evaluation and to determine the number of clusters in this area: water (c_1), deciduous forest (c_2), pasture/hay (c_3), crops (c_4), grains (c_5) woody wetland (c_6) and emergent wetland (c_7), i.e.:

$$\Omega = \{c_1, c_2, c_3, c_4, c_5, c_6, c_7\}$$

We then extract rules from the elevation data where the range of elevations (in metres) for each cluster is calculated as shown in Table 1 (column 2). The conditional probabilities for a sample pixel (p) using SSC are also shown in Table 1 (SSC column).

Table 1. The range of elevation for the selected classes and corresponding SSC probabilities

Clusters	Range of elevation (m)	SSC
c_1	452-459	0.128
c_2	452-478	0.349
c_3	455-478	0.027
c_4	454-490	0.072
c_5	467-481	0.010
c_6	452-472	0.383
c_7	452-477	0.0311

From these ranges we generate a number of rules to either verify the possibility of assigning a pixel to a particular cluster (positive rules) or to reduce the possibility of assigning the pixel to this cluster (negative rules) (or even to change a pixel assignment). A rule of a particular cluster is fired when an elevation is located within the range of elevations of this cluster. For example, we can define a ‘negative’ rule relating the elevation of a pixel e_p , to the possibility of it being classified as mixed forest (c_4):

Rule 1: if ($e_p < l$) OR ($e_p > u$) then class $\neq \{c_4\}$, with confidence interval = [0.95, 1], where l and u are the lower and upper bound for the elevation distribution for c_4 (697-1315). The mass functions for this source of evidence can be expressed as:

$$m\{\Omega \setminus c_4\} = 0.95 \text{ and } m\{\Omega\} = 0.05.$$

where $m\{\Omega \setminus c_4\}$ represents all classes except c_4 . Here we attach most of the mass (0.95) to $m\{\Omega \setminus c_4\}$ and the remainder (0.05) is spread over all classes including c_4 .

Our aim is to incorporate elevation knowledge in a form of such rule(s) into the classification output. Therefore, for pixel p two source of evidence (mass functions) are available for combination. The first piece of evidence is the conditional probabilities resulting from semi-supervised classification which can be formed as masses: $m\{c_1\} = 0.128$, $m\{c_2\} = 0.349\dots$ $m\{c_7\} = 0.0311$. The second source of evidence is the contextual knowledge associated with “elevation” resulting from applying the previous rules. Dempster’s law of combination (equation 2) can then be used to fuse the two masses to get new modified masses (probabilities). Although this approach has shown accuracy improvements in most cases, in some instances the accuracy can get worse. In the following section we show how the rules can be refined in order to address this issue.

2.2 Refined Fusion Process

A new approach is proposed for combining elevation data with spectral data within the framework of Dempster-Shafer evidence theory. The same combination of rules as before is used for this approach; however the confidence for generating a mass function is not fixed as is the case previously. This variation comes from considering the actual “distribution” of elevation values for each class rather than simply the “range” of values. Based on this distribution an “elevation weight” is calculated which is used as the confidence for the corresponding class.

Fig. 4 (a) shows the elevation distribution and range of elevations for the selected classes and represents the number of pixels for each elevation in each corresponding class. Fig. 5 (b) is the equivalent elevation weight in which each elevation is weighted based on the number of pixels within the corresponding classes at each elevation.

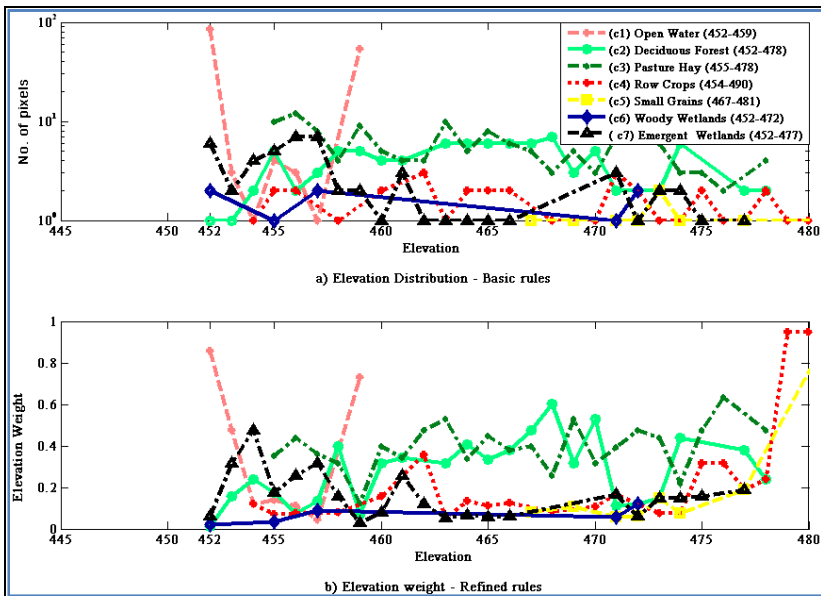


Fig. 4. (a) Elevation distribution (b) Elevation weight

From the elevation ranges (see Table 1) it can be noted that elevations between classes might overlap and therefore different rules attached to these classes are expected to fire for the same elevation. Each of these rules is considered in turn and probabilities modified for those rules which fire. Using the previous approach with the same fixed confidence (0.95) for each rule tends to cause rules to fire that might cancel each other out leading to misclassification.

To illustrate this, we select an elevation that is located within different clusters. The elevation “452” is selected which is located within the range of elevations of c_1 , c_2 , c_6 , and c_7 as shown in Table 1. Therefore, rules attached to these classes are expected to fire. However, the number of pixels at this elevation (452) within these classes are: 85, 1, 2 and 6 respectively as shown in Fig 5 (a). From these values an “elevation weight” can be calculated which represents the confidence for the corresponding cluster. For example, the sum of all pixels attached to the selected elevation is (85+1+2+6 = 94 pixels). Therefore, (for elevation 452), the elevation weight for c_1 is $(85/94) \times 0.95 = 0.859$, for c_2 is $(1/94) \times 0.95 = 0.0101$, for c_6 is $(2/94) \times 0.95 = 0.0202$ and for c_7 is $(6/94) \times 0.95 = 0.0606$ as shown in Fig 5(b). In our previous approach, rules take the same confidence of (0.95). However, we are likely to have a higher confidence for an elevation that is represented by 85 pixels than that represented by 1 pixel.

Table 2 & 3 illustrate the process of using a fixed confidence and a confidence resulting from calculating the elevation weight for each cluster. Therefore, the previous rule (**Rule 1**) can be modified as:

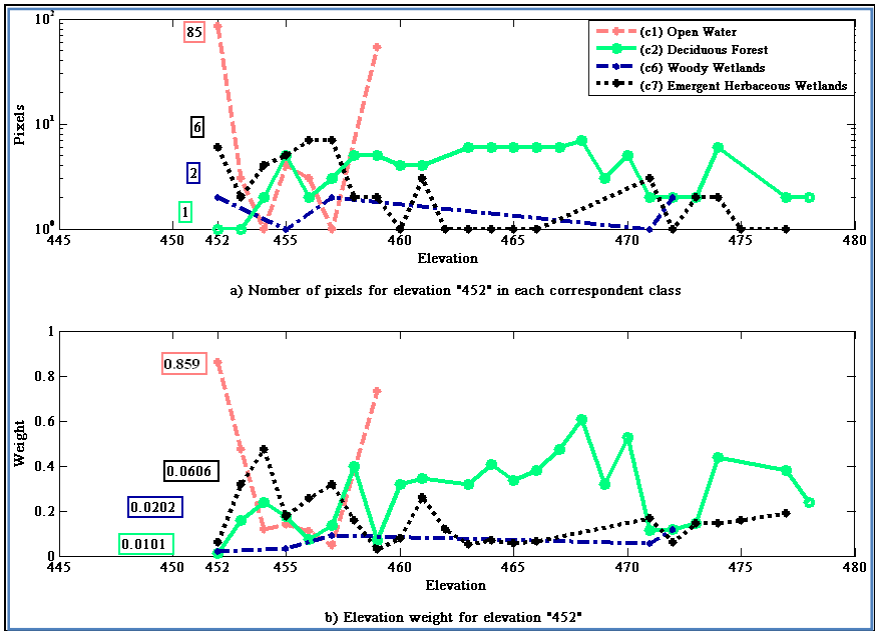


Fig. 5. a) Elevation distribution and b) Elevation weight for elevation “452”

Modified Rule 1: if $(e_p < l)$ OR $(e_p > u)$ then class $\neq \{c_4\}$, with confidence interval = [elevation weight, 1], The mass functions for this source of evidence can be expressed as:

$$m\{\Omega \setminus c_4\} = \text{“elevation_weight” and } m\{\Omega\} = (1 - \text{elevation_weight}).$$

From Table 2, the SSC column represents the original probability resulting from using semi-supervised classification. Rules for classes: c_1, c_2, c_6 and c_7 are expected to fire in sequence. It can be seen that our sample pixel p will be initially assigned to c_6 since it has the highest probability. After firing the rule for c_1 pixel p will be reassigned to c_1 since it has the highest probability now. Eventually, pixel p will be reassigned to c_6 again after firing the rule attached to this cluster. However, the number of pixels at elevation “452” within c_1 is 85, whereas it is 6 for this elevation within c_6 . It is therefore more likely to have more confidence to assign pixel p to c_1 since it has a higher number of pixels than c_6 . Table 3 shows the process of taking the number of pixels into account as an “elevation weight” in our refined approach. It can be seen from Table 3 that pixel p will be assigned to c_1 after firing all the rules in sequence when taking the elevation weight into consideration. This process therefore provides additional confidence when labelling satellite image pixels and also enhances the classification process as a whole.

Table 2. The probabilities resulted for elevation 452 using fixed confidence

Cluster	SSC	Fired classes			
		c_1	c_2	c_6	c_7
c_1	0.349	0.91	0.856	0.729	0.405
c_2	0.128	0.015	0.014	0.012	0.007
c_3	0.027	0.002	0.068	0.059	0.032
c_4	0.072	0.02	0.009	0.156	0.082
c_5	0.01	0.001	0.001	0.001	0.012
c_6	0.383	0.05	0.049	0.042	0.453
c_7	0.031	0.002	0.003	0.001	0.009

Table 3. The probabilities resulted for elevation 452 using “elevation weight” confidence

Cluster	SSC	Fired classes			
		c_1	c_2	c_6	c_7
c_1	0.349	0.469	0.462	0.45	0.449
c_2	0.128	0.096	0.094	0.091	0.092
c_3	0.027	0.026	0.04	0.04	0.029
c_4	0.072	0.059	0.059	0.093	0.084
c_5	0.01	0.009	0.009	0.009	0.008
c_6	0.383	0.328	0.323	0.307	0.319
c_7	0.031	0.013	0.013	0.01	0.019

Table 4. The confusion matrices along with the individual class accuracies for the three approaches

SSC only									SSC with knowledge - Basic rules									SSC with knowledge - Refined rules								
Ref Class.	C1	C2	C3	C4	C5	C6	C7	Row total	Ref Class.	C1	C2	C3	C4	C5	C6	C7	Row total	Ref Class.	C1	C2	C3	C4	C5	C6	C7	Row total
C1	157	3	0	0	0	1	5	166	C1	111	4	0	0	0	1	5	121	C1	163	3	0	0	0	1	3	170
C2	0	46	11	5	0	1	10	73	C2	2	41	11	5	0	1	10	70	C2	0	46	11	5	0	0	5	67
C3	1	11	74	9	1	0	3	99	C3	1	13	74	9	1	0	3	101	C3	1	11	74	8	1	0	3	98
C4	0	2	8	27	0	0	9	46	C4	0	8	8	27	0	0	9	52	C4	0	2	8	32	0	0	5	47
C5	0	17	40	2	6	1	7	73	C5	0	15	40	2	6	1	7	71	C5	0	17	40	2	6	1	4	70
C6	0	20	9	6	2	4	5	46	C6	0	18	9	6	2	4	5	44	C6	0	20	9	4	2	5	5	45
C7	8	4	8	5	0	0	21	46	C7	52	4	8	5	0	0	21	90	C7	2	4	8	3	0	0	35	52
Col total	166	103	150	54	9	7	60	549	Col total	166	103	150	54	9	7	60	549	Col total	166	103	150	54	9	7	60	549

Class	C1	C2	C3	C4	C5	C6	C7	Overall accuracy (%)
SSC only (%)	94.57	44.66	49.33	50.0	66.66	57.14	35.0	61.02
SSC - Basic rules (%)	66.86	39.8	49.33	50.0	66.66	57.14	35.0	51.37
SSC - Refined rules (%)	98.19	44.66	49.33	59.25	66.66	71.42	58.33	65.76

3 Results

The previous study area (Nome) was selected for evaluation. Fig. 6 shows the classified image resulting from performing the three approaches (SSC, SSC-basic rules and SSC-refined rules). The so-called confusion matrix [13] was used for evaluation and the confusion matrices for the three approaches along with individual class accuracies are shown in Table 4. It can be seen from this figure that the overall accuracy when performing SSC only, SSC with knowledge with basic rules and SSC with knowledge with refined rules is 61.02%, 51.37% and 65.75% respectively. These percentages show that using the refined rules has enhanced the classification process as whole. In addition, the individual class accuracies have also improved using this approach.

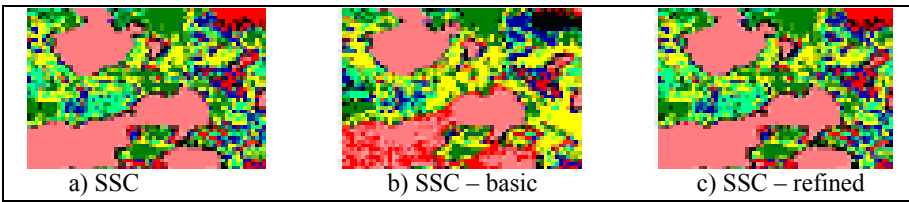


Fig. 6. SSC only and SSC with knowledge – basic and refined rules

Although the overall accuracy may seem not very high, the example has been deliberately chosen to highlight the problem of misclassification as an example in which the accuracy reduced when we applied the basic rules. Also, we have applied our refined approach to many example images and have achieved improved accuracy results in the majority of cases. In addition, we have used our approach with supervised maximum likelihood classification and model-based clustering and in both cases we obtained improved accuracy results. However, our new approach has solved the problem of accuracy degradation using the basic rules because we can control more precisely the manner in which the contextual data is used in the classification process.

4 Conclusions and Future Work

This paper presents a modified approach to enhance satellite image classification by using refined production rules to include contextual data within the classification process. In particular we have demonstrated how Dempster-Shafer's theory of evidence can be used to combine two sources of evidence for the classification of pixel data. A knowledge base is built using specified production rules with different degrees of confidence based on the number of pixels at a particular elevation for each class. Dempster's law of combination is used to combine the two sources of evidence resulting from traditional algorithms and elevation knowledge. A real image was used for illustration showing how using basic rules (fixed confidence) can cause the rules to cancel each other when fired for the same elevation. We then demonstrated how the use of our refined rules can improve the overall classification results.

References

1. Momani, B.M., Morrow, P.J., McClean, S.I.: Using Dempster-Shafer to incorporate knowledge into satellite image classification. *Artif. Intell. Rev.* 25, 161–178 (2006)
2. Dempster, A.P., Laird, N.M., Rubin, D.B.: Maximum likelihood from incomplete data via the EM algorithm. *Royal Statistical Soc. B* 39, 1–39 (1977)
3. Vatsavai, R.R., Shekhar, S., Burk, T.E.: A Semi-Supervised Learning Method for Remote Sensing Data Mining. In: *Proc. ICTAI*, pp. 207–211 (2005)
4. Benkhalifa, M., Bensaid, A., Mouradi, A.: Text categorization using the semi-supervised fuzzy c-means algorithm. *NAFIPS*, 561–565 (June 1999)
5. Huimin, G., Xutao, D., Bastola, D., Ali, H.: On clustering biological data using unsupervised and semi-supervised message passing. *BIBE*, 294–298 (2005)
6. Dempster, A.P.: A Generalisation of Bayesian Inference. *J. of the Royal Statistical Society B* 30, 205–247 (1968)
7. Guan, J.W., Bell, D.A.: *Evidence Theory and its Applications*. Elsevier Science Inc., New York (1991)
8. McClean, S.I., Scotney, B.W.: Using Evidence Theory for the Integration of Distributed Databases. *Int. Journal of Intelligent Systems* 12, 763–776 (1997)
9. Le Héegar-Masclé, S., Richard, D., Ottl'é, C.: Multi-scale data fusion using Dempster-Shafer evidence theory. In: *ICAE*, vol. 10, pp. 9–22 (2003)
10. Yager, R.R., Engemann, K.J., Filev, D.P.: On the Concept of Immediate Probabilities. *Int. Journal of Intelligent Systems* 10, 373–397 (1995)
11. Muralikrishna, I.V.: Image Classification and Performance Evaluation of IRS IC LISS – III Data. In: *IEEE Conference on Geoscience and Remote Sensing*, vol. 4(S), pp. 1772–1774 (1997)
12. The US Geological Survey, <http://www.seamless.usgs.gov> (cited January 10, 2011)
13. Congalton, R.G.: Accuracy assessment and validation of remotely sensed and other spatial information. *Int. Journal of Wildland Fire* 10, 321–328 (2001)

Using Grid Based Feature Localization for Fast Image Matching

Daniel Fleck and Zoran Duric

Department of Computer Science, American University, Washington DC 20016, USA
fleck@american.edu, zduric@cs.gmu.edu

Abstract. This paper presents a new model fitting approach to classify tentative feature matches as inliers or outliers during wide baseline image matching. The results show this approach increases the efficiency over traditional approaches (e.g. RANSAC) and other recently published approaches. During wide baseline image matching a feature matching algorithm generates a set of tentative matches. Our approach then classifies matches as inliers or outliers by determining if the matches are consistent with an affine model. In image pairs related by an affine transformation the ratios of areas of corresponding shapes is invariant. Our approach uses this invariant by sampling matches in a local region. Triangles are then formed from the matches and the ratios of areas of corresponding triangles are computed. If the resulting ratios of areas are consistent, then the sampled matches are classified as inliers. The resulting reduced inlier set is then processed through a model fitting step to generate the final set of inliers. In this paper we present experimental results comparing our approach to traditional model fitting and other affine based approaches. The results show the new method maintains the accuracy of other approaches while significantly increasing the efficiency of wide baseline matching for planar scenes.

1 Introduction

The goal of image matching is to determine if one image matches all or part of another image. This is a fundamental operation for many tasks in computer vision, such as model building, surveillance, location recognition, object detection and many others. In this paper we present a new approach to image matching that increases the efficiency over previous approaches by using affine invariants combined with grid-based localization of features.

In a recent review of matching algorithms conducted by Mikolajczyk and Schmid [1] the algorithms typically have four phases. The first phase detects features in the image. A feature descriptor is created in the second phase to describe each feature. Descriptors are then matched pairwise between the images. Because descriptors describe a small region around a feature, they frequently are incorrectly matched with other small regions that look similar. For example, in a building many windows look the same, thus a window in one image may be matched to multiple windows in another image. Thus, the fourth phase is to filter

out the incorrect matches generated in the third phase by fitting a transformation model to matches and removing any matches that do not fit the transformation. This final model fitting phase is the focus of this paper.

Typical model fitting algorithms generate a model and test it many times until a suitable model is found. In this paper we describe a new approach that uses affine invariant properties to detect and remove incorrect matches before applying standard model fitting approaches. We show that removing incorrect matches early greatly increases the overall efficiency of the matching process. The research presented improves upon earlier work by using a grid-based localization method. We provide results showing the new method is more efficient than recent model fitting algorithms and previous methods using affine invariants.

The remainder of this paper is organized as follows. Section 2 describes other model fitting algorithms which we evaluate. Section 3 describes the new grid-based affine algorithm. Section 4 presents experimental results and Section 5 concludes the paper.

2 Related Work

Matching algorithms typically produce both true positives and false positive matches. With feature based matching algorithms this is unavoidable because a local feature (e.g. a corner of a window) may look exactly like another corner somewhere else in the image. Thus, duplicate features can cause false positive matches when only comparing local neighborhoods. Matching errors also may occur.

Robust methods recover some of the benefits of using global matching algorithms while maintaining the advantage of feature matching algorithms. After a feature matching algorithm has generated a set of putative matches, a robust method will attempt to filter out incorrect matches based on the global properties of the matches. The most popular of these methods is the Random Sample Consensus (RANSAC) method [2].

RANSAC starts by assuming some transformation model (typically affine or perspective) exists between the two images. The minimum number of matches needed to instantiate the model are then randomly selected from the original set of putative matches. From the sample matches a model is created (M). Using M all points P_1 in one image are projected into the second image as P' using Eq. 1. To determine if a match is an inlier when matching images the reprojection error is typically used. Reprojection error (R_{err}) is computed as the Euclidean distance between the matched point P_2 and the corresponding point predicted by the model P'_2 (Eq. 2). Once computed the given point is classified as an inlier or outlier based on this distance (see Eq. 3).

$$P'_2 = M \times P_1 \tag{1}$$

$$R_{err} = d(P_2, P'_2) \tag{2}$$

$$inlier = \begin{cases} true & \text{if } R_{err} < T \\ false & \text{otherwise} \end{cases} \quad (3)$$

Matches predicted with a reprojection below a threshold T are considered inliers and all others are considered outliers. The model with the highest number of inliers is then chosen as the correct image transformation model. By building the model M from a minimal set of matches researchers have shown that RANSAC can determine a correct model with up to 50% outliers in the original set of matches [3].

Due to RANSAC's success, many researchers have worked to improve the efficiency [4,5,6] and the accuracy [7,8,9] of the original approach. In this work we compare our approach to two popular RANSAC variants.

The traditional RANSAC algorithm has a time complexity of $O(mn)$ where m is the number of models evaluated and n is the number of matches checked per model. Many improvements to RANSAC use a heuristic to select probable models which can reduce m . Other approaches reduce n by determining as early as possible that the current model being tested is not correct enabling the algorithm to stop model evaluation. An example of the first approach is described in [5]. In [5] Nister performs a shallow breadth first evaluation of model parameters to determine likely inlier models and then completes the depth first evaluation of only the models with the highest probability of being correct. Nister's system is applied to real-time applications by setting an upper bound on the time available. Using that time available, the number of models to evaluate can be computed. This allows model fitting to be achieved in a real-time system with a trade-off of model accuracy. An example of the second approach is described in [4]. Chum, et. al. use a randomized per-evaluation $T_{d,d}$ test to determine if the current model being evaluated is likely to be a correct model. Using this early exit from the testing process they report an efficiency improvement of an order of magnitude. Recently, Chum improved upon this algorithm with Optimal Randomized RANSAC (RANSAC-SPRT) [10]. RANSAC-SPRT applies Wald's sequential probability ratio test (SPRT) [11] to optimally determine if the current model under evaluation is likely to be a good model. In RANSAC-SPRT Wald's likelihood ratio (Eq. 4) is computed.

$$\lambda_j = \prod_{r=1}^j \frac{p(x_r|H_b)}{p(x_r|H_g)} \quad (4)$$

Where H_b, H_g are a good model and bad model hypothesis, and x_r is 1 if the r^{th} data point is consistent with the model and 0 otherwise. If λ_j is greater than a computed decision threshold, the model is rejected as "bad". The number of points evaluated (j) increases until the model is rejected or all points have been tested. By applying the SPRT algorithm during the model verification, incorrect models can be discarded without verifying all matches in the data set. The resulting algorithm is 2 to 9 times faster than standard RANSAC as reported in [10]. These improvements still require the same number of initial hypothesis as the standard RANSAC algorithm. The efficiency improvements also perform

essentially the same steps as the original RANSAC algorithm, but reduce the number of data points to evaluate through an early evaluation process.

Another popular RANSAC variant is Torr and Zisserman’s Maximum Likelihood Estimation Sample Consensus (MLEM) [7]. In traditional RANSAC to determine if a point is an inlier the reprojection error (R_{err}) is computed as the distance between the predicted location of the matching feature and the observed location of the matching feature. Then the number of points with R_{err} lower than a threshold are considered inliers. In counting a match as either an inlier or an outlier traditional RANSAC doesn’t take advantage of the difference in R_{err} for points within the threshold. Model scoring in MLEM uses the reprojection error distance explicitly by minimizing the maximum likelihood estimate of the reprojection error (e) (given for a single match (i) as Eq. 5).

$$e_i^2 = \sum_{j=1,2} (\hat{x}_i^j - x_i^j)^2 + (\hat{y}_i^j - y_i^j)^2 \quad (5)$$

Thus, points that fit more closely to the model are considered more favorable than points farther from the model. This approach was shown in [7] to find more accurate models than traditional RANSAC. Researchers have built on this approach to create Guided-MLEM [12], NAPSAC [13], MAPSAC [14] and LLN-MLEM [15].

In this paper we compare the accuracy and efficiency of affine model fitting approaches to RANSAC-SPRT and MLEM. At this time no publicly available version of RANSAC-SPRT is available. Thus, experimental results for RANSAC-SPRT use our own implementation of the algorithm. Experimental results for MLEM use the publicly available version from [16]. The results will be presented in section 4.

3 Affine Filtering Using Local Grids

In this work we propose a new way to classify matches as inliers or outliers using affine invariants. Typical classification approaches (as described in section 2) detect inliers by computing a transformation model that maximizes the number of matches that fit the model. Finding the best model requires evaluating a large number of matches during each iteration.

Our approach detects inliers using the property that affine invariant transformations maintain a constant ratio of areas of shapes. We use this property by iteratively sampling four matches in a local region. We then compute the four possible triangles that can be created from the four matches as shown in Fig. 1. If all four triangle pairs produce a consistent ratio of areas the matches are considered inliers.

An affine transformation can model several changes between pairs of images. The transformation has six degrees of freedom including translation in the X and Y directions, rotation, non-isotropic scaling, and shear. The general affine transformation matrix relating image coordinates is shown in Eq. 6.

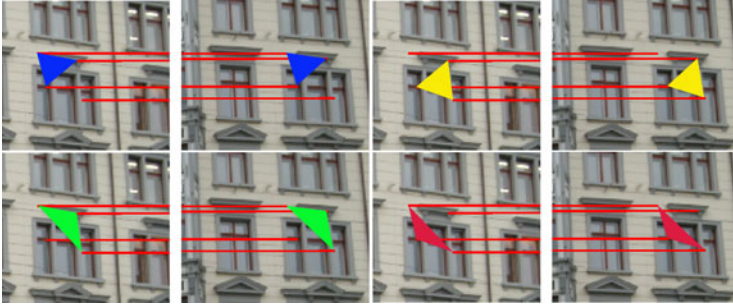


Fig. 1. A single image pair repeated four times showing the four possible triangles created from four pairs of feature matches

$$\begin{pmatrix} x' \\ y' \\ 1 \end{pmatrix} = \begin{pmatrix} a_{11} & a_{12} & t_y \\ a_{21} & a_{22} & t_x \\ 0 & 0 & 1 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} \quad (6)$$

Images related by an affine transformation have invariant properties including parallelism of corresponding lines, ratio of the lengths of corresponding parallel lines and ratios of areas of corresponding shapes [17]. Previous work demonstrated a consistent ratio of areas across an entire image can be used to detect inliers in planar scenes (e.g. images of buildings) [18]. The underlying assumption was that the relationship between image pairs of planar scenes could be approximated by an affine transform. It was shown in [19] that as the perspective distortion grew, this assumption broke down and the algorithm generated unstable results. Further research showed that by selecting features in a local region, the effects of perspective distortion are minimized resulting in better detection of inliers [20]. In [20] features in a local region were selected by computing the Delaunay triangulation [21] of the matches, and then searching the Delaunay graph to pick a random set of neighboring matches. In this work we present results showing the regional affine approach using Delaunay triangulation is less efficient than recent RANSAC improvements (MLESAC and RANSAC-SPRT). We also present a new affine approach using grid-based match selection that retains the set of inliers found by the Delaunay approach, and is more efficient than MLESAC and RANSAC-SPRT.

3.1 Description of Grid-Based Affine Filtering

The grid-based affine method divides the image into a grid. In our work we used grid cells covering 20% of the width by 20% of the image height. Four matches are then randomly selected from a grid cell. The matches are used to generate the four possible triangles. The ratios of areas of corresponding triangles are computed using Eq. 7 where $A_{T_{i,j}}$ is the area of triangle i in image j . The ratios are normalized by dividing them by the maximum ratio as shown in Eq. 8. The

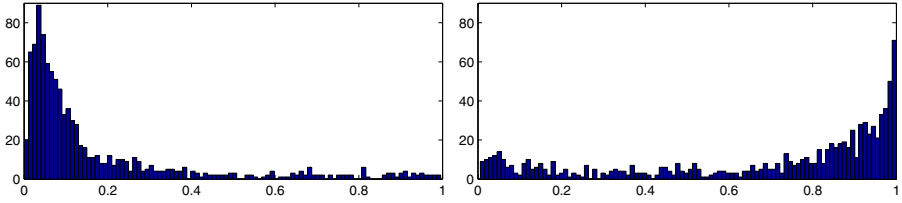


Fig. 2. Histogram of Ratio Differences (R_{diff}). Left: R_{diff} for correct matches only (inliers). Right: R_{diff} for incorrect matches only (outliers).

normalized ratio difference (R_{diff}) is computed as the difference between the maximum and minimum normalized ratios (shown in Eq. 9). If $R_{diff} > \varepsilon$ all matches in the set are discarded as outliers (see Eq. 10) because their ratios of areas are not consistent. This process is repeated up to a set minimum for each grid cell. The resulting inlier set is then processed through MLESAC to create the final set of inliers. In all experiments presented we chose $\varepsilon = 0.10\%$, meaning the difference in area (R_{diff}) must be less than 10% to be considered an inlier. This number was experimentally determined by matching images using traditional methods. The R_{diff} for the inliers and outliers were then plotted as separate histograms as shown in Fig. 2. The figures show that the majority of inliers have an R_{diff} below 20%. In our experiments we chose a more conservative 10% that retains enough inliers to ensure a robust model can be computed.

$$R_1 = \frac{A_{T1,1}}{A_{T1,2}}, \quad R_2 = \frac{A_{T2,1}}{A_{T2,2}}, \quad R_3 = \frac{A_{T3,1}}{A_{T3,2}}, \quad R_4 = \frac{A_{T4,1}}{A_{T4,2}} \quad (7)$$

$$R_{1n} = \frac{R_1}{\max\{R_1, R_2, R_3, R_4\}} \quad R_{2n} = \frac{R_2}{\max\{R_1, R_2, R_3, R_4\}} \quad (8)$$

$$R_{3n} = \frac{R_3}{\max\{R_1, R_2, R_3, R_4\}} \quad R_{4n} = \frac{R_4}{\max\{R_1, R_2, R_3, R_4\}}$$

$$R_{diff} = \max\{R_{1n}, R_{2n}, R_{3n}, R_{4n}\} - \min\{R_{1n}, R_{2n}, R_{3n}, R_{4n}\} \quad (9)$$

$$R_{diff} = \begin{cases} < \varepsilon & \text{inliers} \\ \geq \varepsilon & \text{outliers} \end{cases} \quad (10)$$

This grid-based process is similar to, but much more efficient than, the Delaunay-based regional affine approach presented in [20]. Constructing the Delaunay triangulation is efficient with a time complexity of $O(n \log n)$ [22]. However, to determine the other features in the local region requires repetitively searching the Delaunay graph which is inefficient. In the grid based approach, creating the grid has a constant time complexity, and adding each feature into it’s respective grid cell is an $O(n)$ operation. The grid based affine method is summarized as Alg. 11.

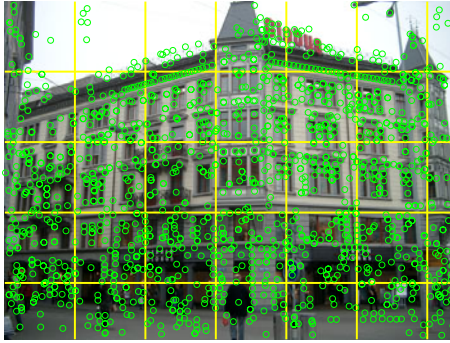


Fig. 3. Sample image showing grid structure imposed by grid-based affine method

Algorithm 1: Grid-based affine method for inlier detection

Input: a set of tentative matches

Output: inliers from the tentative matches

foreach (*match in allMatches*) **do**

Add match into grid cell based on location

end

foreach (*row in grid*) **do**

foreach (*col in grid*) **do**

1. Choose random subsets of four matches in the grid square
2. Compute ratio of areas for subsets
3. Compute normalized ratios (R_n)
4. Compute R_{diff} as the difference between the maximum and minimum normalized ratios
5. Label sets as inliers where $R_{diff} < \varepsilon$

end

end

4 Experimental Results

In this section we present results using real image pairs from the publicly available Zurich Building Database [23] to demonstrate the accuracy and efficiency of the affine algorithms. The original Delaunay based regional affine algorithm and the new grid-based algorithm are compared to baseline RANSAC and the recently published improvements to RANSAC: Maximum Likelihood Estimation Sample Consensus (MLEM) [14] and Optimal Randomized RANSAC (RANSAC-SPRT) [10].

In the experiments presented all code is run as Matlab functions. This provides a valid comparison of the efficiency of each algorithm. The experimental results presented use a publicly available version RANSAC from Kovesi [24] and of MLEM from [16]. In the available MLEM code there was no stopping

criteria, however one was described in the paper. To provide a fair comparison we modified the available implementation to stop when conditions were reached as described in Torr's paper. This change greatly increased the efficiency of the implementation. Due to there being no publicly available Matlab version of RANSAC-SPRT, we have implemented the algorithm based on Chum and Matas' description in [10].

Sample image pairs from the database are shown in Fig. 4.



Fig. 4. Sample image pairs used in tests

4.1 Test Methodology

Lowe's SIFT [25] algorithm generated the tentative matches used as input to the algorithm being tested. Each algorithm being tested was run on the same SIFT matches to generate a set inliers. The resulting inliers were evaluated by applying the normalized direct linear transform algorithm to create a homography transformation [17]. Using the generated transformation the reprojection error was computed using Eq. 2 [26]. All original SIFT matches were then classified as inliers or outliers based on the reprojection error. The results reported label inliers where $R_{err} < 3$ pixels. Experiments with other thresholds were conducted with similar results as those reported here.

Figs. 5, 6, 7, 8 show pairwise comparisons between algorithms. Tests were conducted using all images from the Zurich Building Database comparing the first view to each other view in the database. Each data point represents an image pair. In Figs. 5 and 6 points below the diagonal indicate an image pair where the approach on the X axis took more time than the approach on the Y axis. Fig. 5 shows the results for regional affine approach using Delaunay triangulation. Fig. 6 shows the results for the grid-based affine method. The figures show that while the Delaunay is more efficient than RANSAC and RANSAC-SPRT, in most cases it is not as efficient as MLESAC. Fig. 6 shows that the grid-based affine approach is consistently more efficient than all RANSAC variants tested.

Figs. 7 and 8 show pairwise comparisons of the number of inliers found. Each data point represents an image pair. Points below the diagonal indicate an image

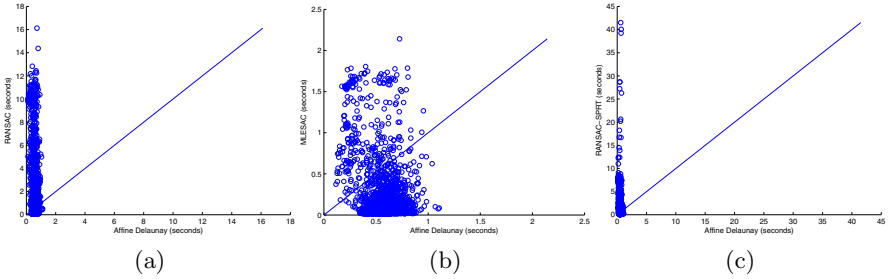


Fig. 5. Pairwise comparison of time to compute inliers using different algorithms. (a) Affine Delaunay versus RANSAC. (b) Affine Delaunay versus MLESAC. (c) Affine Delaunay versus RANSAC-SPRT.

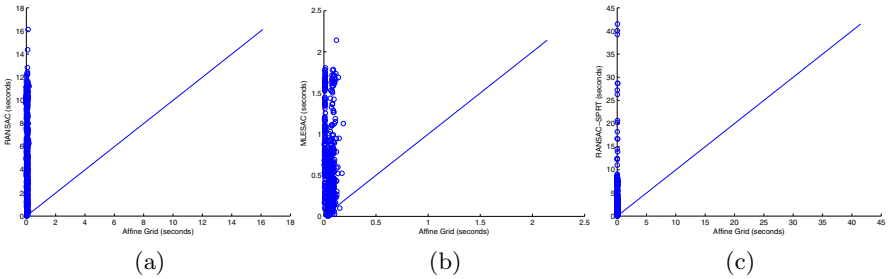


Fig. 6. Pairwise comparison of time to compute inliers using different algorithms. (a) Affine Grid versus RANSAC. (b) Affine Grid versus MLESAC. (c) Affine Grid versus RANSAC-SPRT.

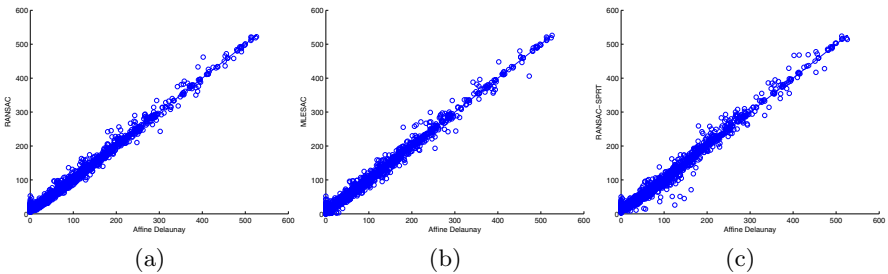


Fig. 7. Pairwise comparison of number of model inliers found. (a) Affine Delaunay versus RANSAC. (b) Affine Delaunay versus MLESAC. (c) Affine Delaunay versus RANSAC-SPRT.

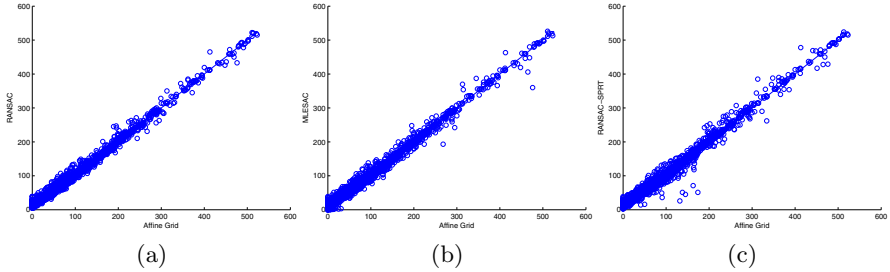


Fig. 8. Pairwise comparison of number of model inliers found. (a) Affine Grid versus RANSAC. (b) Affine Grid versus MLESAC. (c) Affine Grid versus RANSAC-SPRT.

pair where more inliers were found using the approach on the X axis. Both figures show that the affine-based approaches generate very similar inliers as the RANSAC-based approaches. Thus, while maintaining similar accuracy, the grid based affine approach is much more efficient.

5 Conclusion

In this work we have proposed an enhanced method to classify feature correspondences as inliers or outliers. Our approach does not rely on the typical model generation and test approach used by RANSAC-based methods. The grid-based regional affine invariant method samples features from a local region and computes ratios of areas of corresponding triangles. By checking for consistent ratios of areas of the triangles the algorithm can label inliers and outliers. Experiments were performed on a large database of real images. The results show the grid-based affine approach maintains similar accuracy, but is more efficient, than previous affine approaches and recent RANSAC-based model fitting methods.

References

1. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. *IEEE Trans. PAMI* 27(10), 1615–1630 (2005)
2. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. *Commun. ACM* 24(6), 381–395 (1981)
3. Rousseeuw, P., Leroy, A.: *Robust Regression and Outlier Detection*. Wiley, Chichester (1987)
4. Chum, O., Matas, J.: Randomized ransac with td,d test. In: *Proceedings of the 13th British Machine Vision Conference (BMVC)*, pp. 448–457 (2002)
5. Nister, D.: Preemptive ransac for live structure and motion estimation. *MVA* 16(5), 321–329 (2005)
6. Chum, O., Matas, J., Kittler, J.: Locally optimized RANSAC. In: Michaelis, B., Krell, G. (eds.) *DAGM 2003*. LNCS, vol. 2781, pp. 236–243. Springer, Heidelberg (2003)

7. Torr, P.H.S., Zisserman, A.: Mlesac: a new robust estimator with application to estimating image geometry. *Comput. Vis. Image Underst.* 78(1), 138–156 (2000)
8. Tordoff, B., Murray, D.: Guided sampling and consensus for motion estimation. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) *ECCV 2002*. LNCS, vol. 2350, pp. 82–96. Springer, Heidelberg (2002)
9. Wang, H.: Robust adaptive-scale parametric model estimation for computer vision. *IEEE Trans. Pattern Anal. Mach. Intell.* 26(11), 1459–1474 (2004), Senior Member-Suter, David
10. Chum, O., Matas, J.: Optimal randomized ransac. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(8), 1472–1482 (2008)
11. Wald, A.: *Sequential Analysis*. Dover, New York (1947)
12. Tordoff, B.J., Murray, D.W.: Guided-mlesac: Faster image transform estimation by using matching priors. *IEEE Trans. Pattern Anal. Mach. Intell.* 27(10), 1523–1535 (2005)
13. Myatt, D.R., Torr, P.H.S., Nasuto, S.J., Bishop, J.M., Craddock, R.: Napsac: high noise, high dimensional robust estimation. In: *BMVC 2002*, pp. 458–467 (2002)
14. Torr, P.: Bayesian model estimation and selection for epipolar geometry and generic manifold fitting. *IJCV* 50(1), 35–61 (2002)
15. Zhang, L., Rastgar, H., Wang, D., Vincent, A.: Maximum likelihood estimation sample consensus with validation of individual correspondences. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Kuno, Y., Wang, J., Wang, J.-X., Wang, J., Pajarola, R., Lindstrom, P., Hinkenjann, A., Encarnaç o, M.L., Silva, C.T., Coming, D. (eds.) *ISVC 2009*. LNCS, vol. 5875, pp. 447–456. Springer, Heidelberg (2009)
16. Torr, P.: Philip torr’s home page, <http://cms.brookes.ac.uk/staff/PhilipTorr/>
17. Hartley, R.I., Zisserman, A.: *Multiple View Geometry in Computer Vision*, 2nd edn. Cambridge University Press, Cambridge (2004)
18. Fleck, D., Duric, Z.: Affine invariant-based classification of inliers and outliers for image matching. In: Kamel, M., Campilho, A. (eds.) *ICIAR 2009*. LNCS, vol. 5627, pp. 268–277. Springer, Heidelberg (2009)
19. Fleck, D., Duric, Z.: An evaluation of affine invariant-based classification for image matching. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Kuno, Y., Wang, J., Pajarola, R., Lindstrom, P., Hinkenjann, A., Encarnaç o, M.L., Silva, C.T., Coming, D. (eds.) *ISVC 2009*. LNCS, vol. 5876, pp. 417–429. Springer, Heidelberg (2009)
20. Fleck, D., Duric, Z.: Using local affine invariants to improve image matching. In: *International Conference on Pattern Recognition*, pp. 1844–1847 (2010)
21. Barber, C.B., Dobkin, D.P., Huhdanpaa, H.: The quickhull algorithm for convex hulls. *ACM Transactions on Mathematical Software* 22(4), 469–483 (1996)
22. Leach, G.: Improving worst-case optimal delaunay triangulation algorithms. In: *4th Canadian Conference on Computational Geometry*, p. 15 (1992)
23. Griesser, A.: Zurich building database, <http://www.vision.ee.ethz.ch/showroom/zubud/>
24. Kovesi, P.D.: *MATLAB and Octave functions for computer vision and image processing*. School of Computer Science & Software Engineering, The University of Western Australia, [http://www.csse.uwa.edu.au/\\$\sim\\$pk/research/matlabfns/](http://www.csse.uwa.edu.au/\simpk/research/matlabfns/)
25. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60, 91–110 (2004)
26. Ma, Y., Soatto, S., Kosecka, J., Sastry, S.S.: *An Invitation to 3-D Vision: From Images to Geometric Models*. Springer, Heidelberg (2003)

A Hybrid Representation of Imbalanced Points for Two-Layer Matching

Qi Li

Department of Mathematics and Computer Science
Western Kentucky University
qi.li@wku.edu

Abstract. A characteristics of imbalanced points is their localities—an imbalanced point may be contiguous to some other imbalanced points in terms of 8-connectivity. A two-layer scheme was recently proposed for matching imbalanced points based on localities, where the first layer aims to build locality correspondence, and the second layer aims to build point correspondence within corresponding localities. Under the framework of the two-layer matching, we propose a hybrid representation of imbalanced points. Specifically, an imbalanced point in the first layer is represented by a discriminant SIFT-type descriptor, and in the second layer, the imbalanced point is simply represented by a patch-type descriptor (the intensities of its neighborhood). We will justify the rationale of the proposed hybrid representation scheme and show its superiority over non-hybrid representation with experiments.

1 Introduction

Imbalanced points are image points whose first-order derivatives of intensity values can be clustered into two imbalanced classes [10]. Unlike conventional interest points (also called feature points and keypoints) [3][5][2][3][1], an imbalanced point may be contiguous to some other imbalanced points in terms of 8-connectivity [10][8]. This characteristics of imbalanced points is called the *locality property* [9]. One of advantages of localities of imbalanced points is the improved localization accuracy [10] for the higher-level applications, such as stereo correspondence and object recognition.

Based on the locality property, a two-layer scheme was proposed for matching imbalanced points [11], where the first layer (also called the global layer) establishes correspondence between localities, and then the second layer (also called the local layer) refines the locality correspondence to point correspondence. It is intuitive that locality correspondence, with more local information, is more robust with respect to mismatching than conventional point correspondence. In the context of stereo correspondence, the two-layer matching scheme performs a “divide-and-conquer” strategy to address mismatching and imprecise matching separately. (Note that mismatching and imprecise matching are the two main challenges in stereo correspondence [17][5].) An important problem in the two-layer matching scheme is on how to measure the similarity between localities that is equivalent to the problem of measuring similarity of two sets of vectors [11]. Several methods have been proposed to measure similarity of localities, where a similarity measure may or may not be symmetric [8][11][9].

Given an imbalanced point, its local patch, i.e., a window of intensities of its neighboring points, was used as the representation (descriptor) of the point for the two-layer matching scheme [8][19]. The patch-type representation was observed to perform more effectively in the application of the estimation of the fundamental matrix of two stereo images whose baseline is relatively small [9], being compared with SIFT-type representation [12]. But it is also well known that the performance of patch-type representations is poor if images contain significant scaling or affine variations.

The study presented in this paper is motivated by a revisit of the tradeoff between patch-type and SIFT-type representations of interest points, under the framework of the two-layer matching scheme. It is known that that patch-type representations of interest points is generally more sensitive to their localization accuracy than SIFT-type representation. In other words, the patch-type representation of an interest point can be significantly dissimilar to the patch-type representation of a neighboring interest point, while the SIFT-type representation is changed less significantly due to their statistical construction. As specified by Lowe in [12], SIFT can tolerate up to 4-pixel shift due to the design of 4x4 window. This difference leads to the following consequences and tradeoff between the two presentations:

- The patch-type representation tends to perform worse than the SIFT-type representation when imaging variations are significant.
- The patch-type representation tends to contain fewer instances of imprecise matching due to the low tolerance of inaccurate localization.
- Feature extraction methods, such as Principal Component Analysis (PCA) [6][7] and Linear Discriminant Analysis (LDA) [2][16], are expected to be more effective to the SIFT-type presentation than the patch-type representation. Note that these methods assume Gaussian distribution of feature vectors, and the statistical characteristics of SIFT-type representation is more consistent with this assumption.

In this paper, we propose a hybrid representation of imbalanced points for the two-layer matching scheme, based on the locality characteristics of imbalanced points and the tradeoff between the patch-type and SIFT-type representations. Fig. 1 illustrates the basic idea of the proposed hybrid representation. Specifically, we use SIFT-type descriptors of imbalanced points in the first-layer matching (i.e., locality correspondence) in order to tolerate the inaccurate localization of interest points under potentially large imaging variations that in turn aims to reduce mismatching instances. Furthermore, we propose to apply Linear Discriminant Analysis to extract discriminant features of SIFT-type descriptors based on the Fisher criterion [2] that maximizes the intra-locality similarity and minimizes inter-locality similarity of SIFT-type descriptors. In the second-layer matching (i.e., point correspondence within corresponding localities), we use patch-type descriptors of imbalanced points to achieve precise point correspondence.

In the experiment, we test performance of the proposed hybrid representation of imbalanced points in the estimation of fundamental matrices, a well-known problem that is sensitive to mismatching and imprecise matching.

The rest of the paper is organized as follows: In section 2, a hybrid representation of imbalanced points is proposed. Experiments are presented in Section 3. Finally, conclusions are given in Section 4.

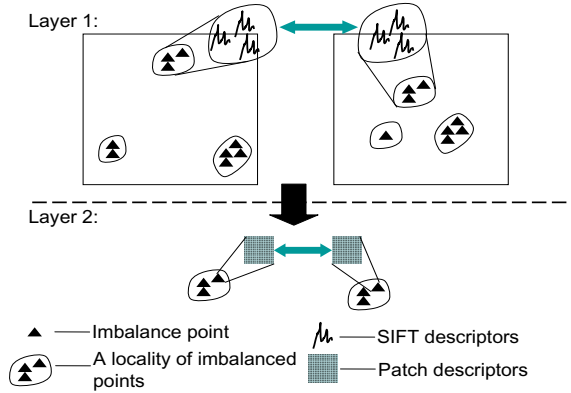


Fig. 1. A hybrid representation of imbalanced points for the two-layer matching scheme

2 A Hybrid Representation for Two-Layer Matching

This section contains three parts. In the first part, we review a similarity measure of two localities based on patch-type descriptors. In the second part, we propose a SIFT-type descriptor of an imbalanced point and its Fisher discriminant representation. In the last part, we propose a hybrid representation scheme that integrates the superiority of patch-type and (discriminant) SIFT-type representations.

In the rest of the paper, we assume that a patch of an imbalanced point has been normalized in terms of its scale and orientation (or more generally affine parameters). It is worth noting that it is not straightforward to assign a characteristic scale and orientation to an imbalanced point without involving non-maximum suppression. But we can still assign a scale and orientation to an imbalanced point by a certain association approach between the imbalanced point and an interest point with known scale and orientation (such as a Lowe’s keypoint [12]). We skip the details on the scale and orientation selection in this paper due to the space constraint.

2.1 Similarity Measure of Localities

Given two images I_1 and I_2 , denote P and P' as sets of imbalanced points of I_1 and I_2 respectively. Note that the cardinality of two localities P and P' , i.e., the numbers of imbalanced points, may be different. Denote the descriptors of imbalanced points in a locality P_i as $D_{P_i} = \{d_p | p \in P_i\}$, where d_p is a descriptor of a point p . We propose the following idea to measure the similarity between two localities P_i and P'_j , more specifically, the similarity between D_{P_i} and $D_{P'_j}$: for each $p \in P_i$, we find its most similar point $p' \in P'_j$. It is possible that two different points p_{i_1} and p_{i_2} are found to correspond to the same p' . So here, we have many-to-one and imprecise point correspondence. Note that many-to-one point correspondence can happen since in certain image variations, such as resolution (scale), two localities associated with the same physical scene region may not have the exact same number of imbalanced points. Also note that imprecise correspondence in global scale is not an issue since it will be refined

in the local scale. Intuitively, the overall within-locality point similarity is high if two localities associated with the same physical scene region.

Assume that D_{P_i} and $D_{P'_j}$ are patch-type descriptors. A similarity measure between D_{P_i} and $D_{P'_j}$ proposed in the previous work [11] is defined as follows:

$$S(D_{P_i}, D_{P'_j}) = \frac{1}{|P_i|} \sum_{p \in P_i} \max_{p' \in P'_j} S(d_p, d_{p'}),$$

where $|\cdot|$ denotes the cardinality of a set. This similarity measure was motivated by the fact that patch-type descriptors are sensitive to localization, and thus the correlation between $p \in P_i$ and $q \in P'_j$ varies significantly.

2.2 Locality Similarity Based on SIFT-Type Descriptors

Since SIFT-type descriptors are robust to localization error of interest points, and the correlation between p and q is expected to be stable with respect to various imaging conditions, we propose the following similarity measure for localities of SIFT-type descriptors:

$$S(D_{P_i}, D_{P'_j}) = \frac{1}{|P_i| \times |P'_j|} \sum_{p \in P_i, p' \in P'_j} S(d_p, d_{p'}),$$

where $|\cdot|$ denotes the cardinality of a set.

In this paper, we will basically follow the construction method of Lowe’s SIFT representation, i.e., the concatenation of orientation histograms of 4×4 sub-blocks subdivided from a patch [12]. But, we propose an alternative option to select an underlying patch to construct a SIFT descriptor to address a subtle difference between an imbalanced point and a scale-invariant interest “point” such as a Lowe’s keypoint [12] or Harris-Laplace “points” [14]). Specifically, a scale-invariant interest “point” may be more strictly called an interest region that can be intuitively interpreted as a region surrounded by edges, either entirely or partially. So, the patch used to construct a SIFT descriptor is generally centered by the point, which is consistent with the surrounded region associated with the point.

However, an imbalanced point itself is not a region. (This is also a reason why it is not straightforward to assign scale and orientation to an imbalanced point as mentioned before.) An imbalanced point usually lies in a border (more precisely a corner) of a region, as illustrated by Fig. 2. So a patch centered by an imbalanced point likely covers a foreground area (i.e., a potential interest region) and a background area. In Fig. 2, the dash circle visualize the scale (i.e., a patch) of an imbalanced point. Note that this patch covers both a foreground area and a background area, where the background area is larger than the foreground area. In a 3D world, a background area tend to vary more significantly with respect to a viewpoint change than a foreground area. Thus, a patch centered by the imbalanced point may not be the best to construct a descriptor invariant to a viewpoint change. We propose to use a patch centered by a point that is drifted from the p along its orientation with the distance of the scale. Fig. 2 illustrates the idea of “drifting a patch”, where the solid circle and solid arrow visualize the scale and orientation of the drifted point.

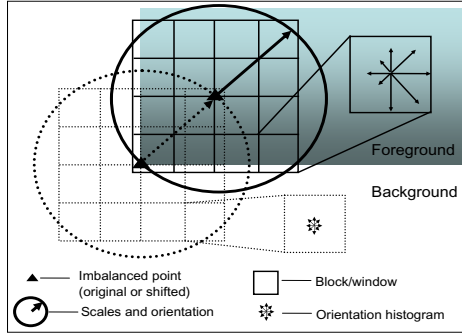


Fig. 2. Illustration of SIFT-type descriptors of imbalanced points

2.3 Discriminant SIFT-Type Descriptors

We propose a discriminant statistic model of the SIFT-type descriptors of imbalanced points based on the locality property of imbalanced points, in the context of stereo correspondence. Here, the sample space for the statistic model includes all descriptors from a single image (one image of a stereo pair). In terms of classification terminology, we consider a locality as *a class*. Specifically, we consider the descriptors of imbalanced points in the same locality a set of instances of a class. With the above assumption that patches have been normalized according to imaging conditions, such as scales and orientations, the variation (uncertainty) among descriptors of imbalanced points in the same locality are mainly caused by localization shift. If descriptors we use are Lowe's SIFTs, then the intra-locality descriptor similarity is expected to be high, i.e., small variations intra-locality due to robustness of orientation histogram with respect to localization inaccuracy.

We construct discriminant descriptors under the Fisher criterion that maximizes similarities of SIFT-type descriptors within the same locality and minimizes the similarities of descriptors between different localities, simultaneously. For convenience, we call the first type of similarity *intra-locality similarity*, and the second type of similarity *inter-locality similarity*. Linear discriminant analysis is a popular feature extraction scheme in classification applications, such as face recognition [2].

PCA-SIFT [7] is a statistic model of descriptors proposed in the context of image retrieval, where Principal Component Analysis (PCA) [6] was applied to local gradient patches (rather than Lowe's histogram of oriented gradients) of keypoints detected from a set of training images of diverse scenes. Theoretically, PCA-SIFT approach requests a sufficiently large number of training images in order to collect a reliable sample space to build a statistic model. Thus, the PCA-SIFT rationale may be not generally effective such as in the applications of stereo correspondence, tracking, etc. The key motivation of the authors introducing PCA-SIFT is the simplicity of the PCA approach, and the reduced dimension to speedup retrieval, although it is assumed that data to model by PCA should satisfy a Gaussian distribution. Recall that our motivation of applying LDA to Lowe's SIFT descriptors is that orientation histograms of imbalanced points within each locality may satisfy Gaussian distribution.

Given an image I , we denote N the number of imbalanced points of I , d_p descriptor of an imbalanced point p , n the dimension of a descriptor, P_i i -th locality of imbalanced points, N_i is the number of imbalanced points in i th locality, and k the number of localities, Denote D_{P_i} a matrix of the descriptors of imbalanced points in the locality P_i . The (local) centroid of descriptors in i -th locality P_i is denoted as $\bar{d}_{P_i} = \frac{1}{N_i} \sum_{p \in P_i} d_p$. The global centroid of descriptors of all imbalanced points in image I is denoted as $\bar{d} = \frac{1}{k} \sum_{i=1}^k \bar{d}_{P_i}$.

The inter-locality distance of descriptors of I can be formulated as

$$\Delta_B = \sum_{i=1}^k \frac{N_i}{N} \|\bar{d}_{P_i} - \bar{d}\|^2, \quad (1)$$

and the intra-locality distance of descriptors of I can be formulated as

$$\Delta_W = \sum_{i=1}^k \sum_{p \in P_i} \|d_p - \bar{d}_{P_i}\|^2, \quad (2)$$

We define inter-locality scatter matrix S_b and intra-locality scatter S_w as follows:

$$\begin{aligned} S_b &= \frac{1}{N} \sum_{i=1}^k N_i (\bar{d}_{P_i} - \bar{d})(\bar{d}_{P_i} - \bar{d})^T = \frac{1}{N} H_b H_b^T, \\ S_w &= \frac{1}{N} \sum_{i=1}^k \sum_{p \in P_i} (d_p - \bar{d}_{P_i})(d_p - \bar{d}_{P_i})^T = \frac{1}{N} H_w H_w^T, \end{aligned} \quad (3)$$

where

$$\begin{aligned} H_b &= [\sqrt{N_1}(\bar{d}_{P_1} - \bar{d}), \dots, \sqrt{N_k}(\bar{d}_{P_k} - \bar{d})] \in R^{n \times k}, \\ H_w &= [D_{P_1} - \bar{d}_{P_1} \cdot e_1, \dots, D_{P_k} - \bar{d}_{P_k} \cdot e_k] \in R^{n \times N}, \end{aligned}$$

$$e_i = (1, \dots, 1) \in R^{1 \times N_i}.$$

LDA is commonly found by solving the trace optimization of the following,

$$G = \arg \max_G \text{trace} \left((G^T S_w G)^{-1} (G^T S_b G) \right), \quad (4)$$

The optimization criterion in (4) is equivalent to the following generalized eigen problem,

$$S_b x = \lambda S_w x, \text{ for } \lambda \neq 0. \quad (5)$$

The solution can be obtained by solving an eigen problem on matrix $S_w^{-1} S_b$. There are at most $k - 1$ non-zero eigenvalues, since the rank of the matrix S_b is bounded by $k - 1$. Therefore, the reduced dimension by LDA is at most $k - 1$. A method to solve the eigen problem is to apply SVD on the scatter matrices.

Algorithm 1 summarizes the steps for the construction of hybrid descriptors of imbalanced points in the context of stereo correspondence.

Algorithm 1. Construction of hybrid descriptors

Input: P - Imbalanced points detected in image I
 P' - Imbalanced points detected in image I'

Output:

Part I: Construct normalized patch-type and SIFT-type descriptors

1. Construct patch-type descriptors D^{patch}
2. Construct SIFT-type descriptors D^{sift}

Part II: Construct Fisher-SIFT

1. Construct inter-locality scatter S_B and intra-locality scatter S_W .
 2. $G \leftarrow \operatorname{argmax}_G \frac{\operatorname{trace}(GS_B G^T)}{\operatorname{trace}(GS_W G^T)}$
 3. $\tilde{D}^{sift} = D^{sift} * G$
 4. $\tilde{D}^{sift, \prime} = D^{sift, \prime} * G$
 5. Normalize each vector in \tilde{D}^{sift} and $\tilde{D}^{sift, \prime}$ as a unit vector
-

3 Experiments

In this section, we will test the performance of proposed hybrid representation in point correspondence and the estimation of the fundamental matrix (i.e., epipolar geometry). Given a set of initial point correspondence, we apply the commonly used linear randomized approach to estimate its fundamental matrix, i.e., 8-point algorithm with RANSAC to estimate the fundamental matrix [4]. Note that RANSAC is a randomize approach to prune mismatches (outliers). The threshold on the residue of a point according to an estimated fundamental matrix is set to 1.5 pixels, and the number of iterations for one round of RANSAC is 500.

We first compare the performance of hybrid descriptors with non-hybrid descriptors (i.e., patch-type descriptors and SIFT-type descriptors). The test images are boat images from INRIA dataset that contains large scale and orientation variations. Fig. 3 shows matched imbalanced points (with scales) and estimated epipolar lines overlaid on input images via three different representations. The number of matching pairs via patch-type descriptors (79 pairs) is fewer than the number of matching pairs via SIFT-type descriptors (98 pairs). This is a common phenomenon in the comparison between patch-type and SIFT-type descriptors due to the higher-tolerance of localization errors of SIFTs than patches. The number of matching pairs via hybrid descriptors (116) is the largest. The numbers of mismatched or inaccurately matched pairs via the three methods are 37, 32, and 21 respectively, which indicates that the ratio of (accurate) matching pairs via hybrid descriptors is also the highest one. Note that the estimated epipolar geometry via hybrid descriptors is the one most consistent with the ground truth.

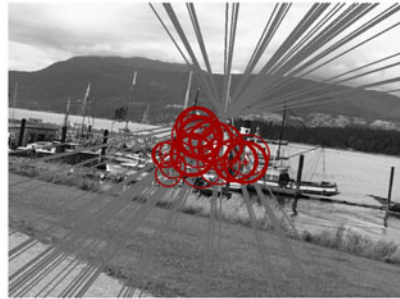
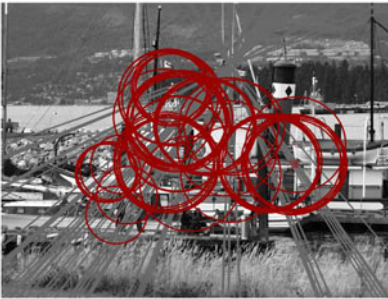
Next, we present a test of hybrid representation of imbalanced points on Middlebury stereo dataset 2006. Middlebury stereo images have been introduced as a ground truth dataset for the study of dense stereo correspondence. Stereo pairs of images are taken by camera translations parallel to the optical plane, and they are further rectified so that all epipolar lines are horizontal. In contrast to previous datasets, dataset 2006



(a) Patch-type descriptors



(b) SIFT-type descriptors



(c) Hybrid descriptors

Fig. 3. Epipolar geometry estimated using (a) patch-type descriptors, (b) SIFT-type descriptors, and (c) hybrid descriptors. The numbers of matching pairs are 79, 98, and 116, respectively. The third result is most consistent with the ground truth.

introduced new challenges on the textures of scenes. One challenge is sparse textures. Fig. 4 lists three examples of tested stereo images, named flowerpot, plastic, and lampshade, respectively. The stereo views of tested images in Fig. 4 are named view 0 and view 6, respectively, in the Middlebury dataset.

Sparse textures bring a challenge to interest point correspondence with the estimation of an epipolar geometry as well as dense correspondence. Note that detecting large

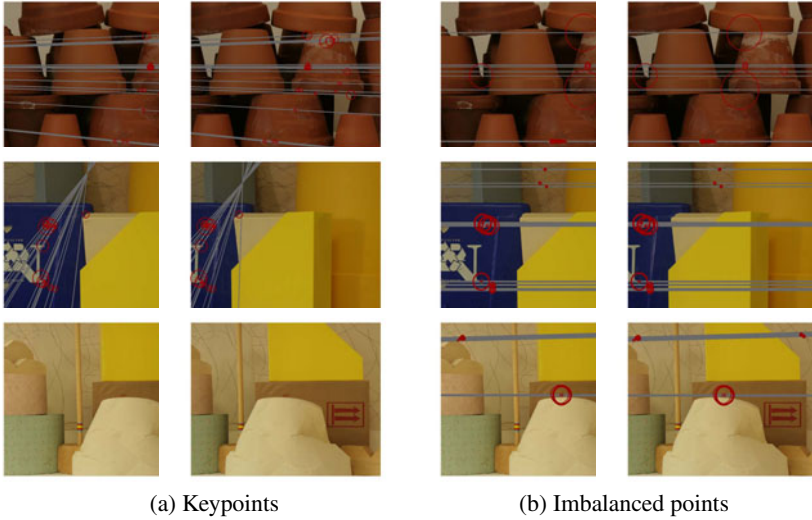


Fig. 4. Test on Middlebury 2006 stereo dataset

numbers of interest points and thus building sufficient initial point correspondence (based on NCC) has been a motivation for the state-of-the-art detectors [12][13]. From Fig. 4 it is clear to observe that an epipolar geometry estimated via imbalanced points with scales is much more consistent with the ground truth than the epipolar geometry estimated via keypoints. For the lampshade images, insufficient matched points are found to estimate a fundamental matrix. The numbers of matched pairs of imbalanced points on these three pairs of images are 31, 73, and 38, respectively. It is worth noting that the numbers of matched pairs of keypoints extracted from these sparsely-textured images are small, and moreover, the scales of matched keypoints are also small. The latter phenomenon may be caused by the following reason—the localization of keypoints of larger scales from sparsely-textured images is less accurate than keypoints of smaller scales, and descriptors of keypoints of larger scales are less distinctive than descriptors of keypoints of smaller scales.

4 Conclusions

In this paper, we propose a hybrid representation of imbalanced points for the two-layer matching scheme, where the first-layer matching is based on discriminant SIFT-type descriptors of imbalanced points, and the second-layer matching is based on patch-type descriptors. Experiments show the effectiveness of the hybrid representation.

Acknowledgment. The work of Q. Li was supported by National Science Foundation Grant IIS-1016668 and the Summer Faculty Scholarship 10-7052 of Western Kentucky University.

References

1. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.J.: Speeded-up robust features (surf). *Computer Vision and Image Understanding* 110(3), 346–359 (2008)
2. Belhumeur, P.N., Hespanha, J., Kriegman, D.J.: Eigenfaces vs. Fisherfaces: Recognition using class specific linear projection. In: Buxton, B.F., Cipolla, R. (eds.) *ECCV 1996. LNCS*, vol. 1064, pp. 45–58. Springer, Heidelberg (1996)
3. Harris, C., Stephens, M.: A combined corner and edge detector. In: *Proc. 4th Alvey Vision Conference*, Manchester, pp. 147–151 (1988)
4. Hartley, R.I.: In defense of the eight-point algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 19(6), 580–593 (1997)
5. Hartley, R.I., Zisserman, A.: *Multiple View Geometry in Computer Vision*. Cambridge University Press, Cambridge (2000)
6. Jolliffe, I.T.: Principle component analysis. *Journal of Educational Psychology* 24, 417–441 (1986)
7. Ke, Y., Sukthankar, R.: Pca-sift: A more distinctive representation for local image descriptors. In: *CVPR (2)*, pp. 506–513 (2004)
8. Li, Q.: Interest points of general imbalance. *IEEE Transactions on Image Processing* 18(11), 2536–2546 (2009)
9. Li, Q., Xing, G.: New similarity measures of localities for a two-layer matching scheme and estimation of fundamental matrices. *Neurocomputing* 73(16–18), 3114–3122 (2010)
10. Li, Q., Ye, J., Kambhamettu, C.: Interest point detection using imbalance oriented selection. *Pattern Recognition* 41(2), 672–688 (2008)
11. Li, Q., Xia, Z., Tao, D.: A global-to-local scheme for imbalanced point matching. In: *IEEE International Conference on Image Processing (ICIP 2009)*, pp. 2117–2120 (2009)
12. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
13. Mikolajczyk, K., Schmid, C.: Scale & affine invariant interest point detectors. *International Journal of Computer Vision* 60(1), 63–86 (2004)
14. Mikolajczyk, K., Tuytelaars, T., Schmid, C., Zisserman, A., Matas, J., Schaffalitzky, F., Kadir, T., Van Gool, L.J.: A comparison of affine region detectors. *International Journal of Computer Vision* 65(1–2), 43–72 (2005)
15. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of interest point detectors. *International Journal of Computer Vision* 37(2), 151–172 (2000)
16. Swets, D.L., Weng, J.J.: Using discriminant eigenfeatures for image retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 18(8), 831–836 (1996)
17. Zhang, Z., Deriche, R., Faugeras, O.D., Luong, Q.-T.: A robust technique for matching two uncalibrated images through the recovery of the unknown epipolar geometry. *Artificial Intelligence* 78(1–2), 87–119 (1995)

Wide-Baseline Correspondence from Locally Affine Invariant Contour Matching

Zhaozhong Wang and Lei Wang

Beihang University
Image Processing Center
Beijing 100191, China

Abstract. This paper proposes an affine invariant contour description for contour matching, applicable to wide-baseline stereo correspondence. The contours to be matched can be either object edges or region boundaries. The contour descriptor is constructed locally using matrix theory and is invariant to affine transformations, which approximate perspective transformations in wide-baseline imaging. Contour similarity is measured in terms of the descriptor to establish initial correspondence, then new constraints of grouping, ordering and consistency for contour matching are introduced to cooperate with the epipolar constraint to reject outliers. Experiments using real-world images validate that the proposed method results in more accurate stereo correspondence for clutter scenes with large depth of field than point-based stereo matching algorithms.

1 Introduction

Wide-baseline stereo matching is a challenging problem in computer vision and many excellent algorithms have been proposed in this field. The method proposed in this paper uses contour cue for wide-baseline stereo correspondence. We first extract contours using state-of-the-art contour detection methods, then construct affine invariant descriptor to characterize the contours. The contours between different views are finally matched in terms of the descriptor and matching outliers are rejected using the epipolar constraint combined with new constraints suitable for contour matching. The contour-based matching is robust to object occlusion between view changes and generates more accurate stereo correspondence.

The proposed method is different from point-based stereo matching. A number of stereo matching methods based upon feature points have been studied in the literature, for example the SIFT feature [12] and affine invariant features [15], which are invariant to image transformations. The proposed method is also distinguished from the approaches of affine invariant regions [19] for stereo matching since they do not use the contour information.

Recently, the contour-based stereo methods have a great development. The first research direction is mainly on improving the speed and accuracy. For example, the work of [16] focus on accelerating the edge-based stereo algorithm

through the reduction of search space; the researches in [9,17] apply belief propagation on edge images to get more accurate stereo correspondence. When scene structures do not lie in or near the frontal parallel plane, Zucker [11,10] proposed a differential geometrical model which relates the structures in the left and right images using the Frenet geometry of space curves.

The second improvement of contour-based stereo matching mainly focus on wide-baseline correspondence. The difficulty that edges of left and right images may be inconsistent due to viewpoint changing would make the matching impossible. To handle this problem, Meltzer and Soatto [14] proposed a bipartite edge descriptor for matching and used ordering constraint to improve the result. The bipartite descriptor is formed by image intensity and gradient information, and two descriptors are required for one occluding edge. Our method in this paper is different from [14] because we use purely geometric information of contours to form their descriptions, so it is simpler in theory. In addition to the ordering constraint, we also use grouping and consistency constraints to refine matching results.

We shall propose the theory of locally affine invariant contour description in Section 2, then describe the contour matching scheme in Section 3. Experimental results of our algorithm with comparisons to other methods on wide-baseline stereo are shown and discussed in Section 4. Our conclusion is presented in the last section.

2 Affine Invariant Contour Description

The proposed algorithm requires detected contours for matching. Detailed schemes for contour detection will be proposed later; and we assume here that we have obtained image contours. Now we describe the theory of contour invariance for stereo correspondence. For wide-baseline matching, a contour description should be invariant to perspective transformations and viewpoint changes. There are work for viewpoint invariant features [20], but for simplicity and efficiency, we shall use the affine invariance to approximate the view invariance. This is similar to many point-based features [12,15] used for correspondence. Our method uses geometric information along contours to describe them and is derived from matrix theory; this is different from most point descriptors and contour descriptors [14] formed by image intensity.

A contour in a 2D image plane can be represented by an ordered point set $\mathcal{C} = \{\mathbf{x}_1, \dots, \mathbf{x}_m\} \subset \mathbb{R}^2$. For a point $\mathbf{x}_i \in \mathcal{C}$, we take its neighboring n points $\mathbf{x}_k \in \mathcal{C}$, $k = i - s, \dots, i + s$ to characterize \mathbf{x}_i , where $n = 2s + 1$ and s an integer (we usually take $s = 15$). Then we construct the following *configuration matrix*

$$X_i = [\mathbf{x}_{i-s}, \dots, \mathbf{x}_i, \dots, \mathbf{x}_{i+s}]^T \in \mathbb{R}^{n \times 2}, \quad (1)$$

and assume that it is of full rank. Under the wide-baseline imaging, the configuration X_i might be transformed perspectively, we approximate this local transform of points using an affine transformation of the form

$$[Y_j \ \mathbf{1}_n] = [X_i \ \mathbf{1}_n] \begin{bmatrix} A & \mathbf{0} \\ \mathbf{t}^T & 1 \end{bmatrix}, \quad (2)$$

where Y_j represents the configuration of a point \mathbf{y}_j which is the correspondence of \mathbf{x}_i , $\mathbf{1}_n := [1, \dots, 1]^T \in \mathbb{R}^n$, A is a 2×2 nonsingular matrix representing affine transformations like rotation, scaling and shearing, and $\mathbf{t} \in \mathbb{R}^2$ is a translation vector. We can first remove the effect of translation by centering the configurations using a formula like

$$\check{X}_i = \left(I - \frac{1}{n} \mathbf{1}_n \mathbf{1}_n^T \right) X_i,$$

then Eq. (2) reduces to

$$\check{Y}_j = \check{X}_i A. \tag{3}$$

We now need to establish affine invariant descriptors for the configurations \check{Y}_j and \check{X}_i . To this end, we compute orthonormal matrices Ω_j and Θ_i from \check{Y}_j and \check{X}_i respectively, where $\Omega_j, \Theta_i \in \mathbb{R}^{n \times 2}$ and $\Omega_j^T \Omega_j = \Theta_i^T \Theta_i = I$. Then we have

$$\Omega_j = \Theta_i O, \tag{4}$$

where $O \in \mathbb{R}^{2 \times 2}$ is an orthogonal matrix. There are several ways to prove this relation, the following is a simple one: Eq. (3) implies that the two matrices \check{Y}_j and \check{X}_i have the same column space [8], thus the columns of Ω_j and those of Θ_i form two orthonormal bases for the same space. This indicates that the two bases are related by an orthogonal transformation like Eq. (4). Zuliani et al. [21] proposed a result similar to Eq. (4); their result is only applicable to closed contours, while Eq. (4) is applicable to both closed and open contours. In addition, their work used a shape matrix-based descriptor for contour matching, which is more complex than our descriptor, as proposed below.

By orthonormalizing the configuration matrices, the transformation between two configurations reduces from an affine A to an orthogonal O , while the latter has many useful properties. For example, an orthogonal transformation does not change the norm of a vector. Let

$$\Theta_i = [\mathbf{q}_{i-s}, \dots, \mathbf{q}_{i+s}]^T,$$

where $\mathbf{q}_k^T \in \mathbb{R}^2$, $k = i - s, \dots, i + s$, are row vectors of Θ_i . These row vectors \mathbf{q}_k^T can be viewed as orthonormalized versions of the original row vectors \mathbf{x}_k^T (i.e., the coordinates of contour points) in Eq. (1). We take the norms of the row vectors \mathbf{q}_k^T to form a new vector

$$\mathbf{w}_{X_i} := [\|\mathbf{q}_{i-s}\|_2, \dots, \|\mathbf{q}_{i+s}\|_2]^T \in \mathbb{R}^n. \tag{5}$$

Note that \mathbf{w}_{X_i} is equal to the square root of the diagonal vector of the matrix $\Theta_i \Theta_i^T$. From Eq. (4) we know that $\Omega_j \Omega_j^T = \Theta_i O O^T \Theta_i^T = \Theta_i \Theta_i^T$, thus the vector \mathbf{w}_{X_i} is invariant to any orthogonal transformation O and in turn invariant to any affine transformation A . We use the vector \mathbf{w}_{X_i} as an affine invariant descriptor of the contour point $\mathbf{x}_i \in \mathbb{R}^2$; it is also the descriptor of the local contour

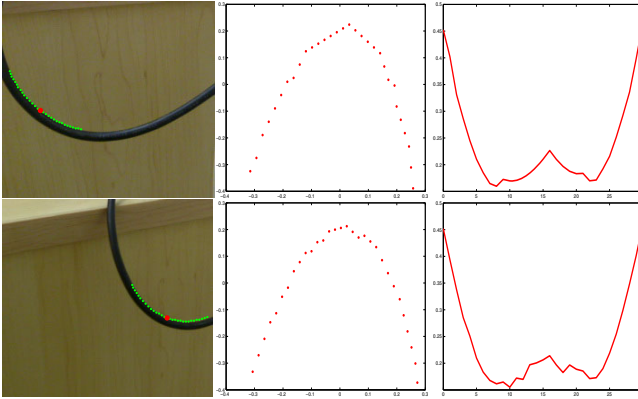


Fig. 1. Illustrating the proposed contour descriptor. The left two images contain contours to be matched, where the red dots are two matched points \mathbf{x}_i and \mathbf{y}_j and the green points near them are those to form local configurations X_i and Y_j . The middle two plots show the orthonormal matrices Θ_i and Ω_j , where each row vector of them is plotted as a dot in \mathbb{R}^2 plane. The right two plots illustrate the descriptor vectors \mathbf{w}_{X_i} and \mathbf{w}_{Y_j} .

configuration X_i . The descriptor will be measured in terms of vector norms to determine if two contour points are matched, as shown in Section 3.

The orthonormal matrix Θ_i can be computed from the centered configuration matrix \check{X}_i by fast numerical methods [8] like the QR factorization or the Gram-Schmidt orthogonalization. Then the descriptor \mathbf{w}_{X_i} is constructed by the norms of row vectors of Θ_i . This pipeline is very simple and efficient. Fig. 1 illustrates the contour descriptor.

There are two remarks here. First, the similarity of two contour configurations under affine transformation can be measured by a distance between two subspaces [2] as well. This would be more robust, but also more time consuming than the proposed vector of descriptor since more computations like the singular value decomposition are required. Second, there may exist a reverse ordering of contour points due to large degrees of in-plane rotation. In this case Eq. (2) does not hold directly, and the reverse of point order should be taken into account. For typical wide-baseline stereo imaging, the reverse of contour points is not critical, so we omit its effect in this paper.

3 Contour Correspondence

The proposed affine invariant descriptor approximates the perspective transformations on a local segment of contour; this makes the descriptor suitable for wide-baseline stereo. In this section we use the descriptor to match contour points between two views. Then we use new global constraints such as grouping and consistency to reject matching outliers.

3.1 Initial Matching from Contour Descriptor

We assume in general that an image to be matched contains multiple contours \mathcal{C}_α , $\alpha = 1, 2, \dots$, and another image also contains multiple contours \mathcal{S}_β , $\beta = 1, 2, \dots$. The number of contours and the number of points on each contour can both be different for the two images. We aim to establish a correspondence of points between the two contour sets

$$\cup \mathcal{C} := \bigcup_\alpha \mathcal{C}_\alpha, \quad \cup \mathcal{S} := \bigcup_\beta \mathcal{S}_\beta,$$

i.e., look for a point mapping $f : \cup \mathcal{C} \rightarrow \cup \mathcal{S}$.

For each point $\mathbf{x}_i \in \mathcal{C}_\alpha \subset \cup \mathcal{C}$ we pick up n points (typically $n = 31$) in its neighborhood along the contour \mathcal{C}_α to form a local configuration $X_i \in \mathbb{R}^{n \times 2}$, then compute the local descriptor $\mathbf{w}_{X_i} \in \mathbb{R}^n$. Similarly, for each point $\mathbf{y}_j \in \mathcal{S}_\beta \subset \cup \mathcal{S}$ we compute its descriptor $\mathbf{w}_{Y_j} \in \mathbb{R}^n$. To determine if the two points \mathbf{x}_i and \mathbf{y}_j would be matched to each other, we use the following simple measure

$$\gamma(\mathbf{x}_i, \mathbf{y}_j) = \|\mathbf{w}_{X_i} - \mathbf{w}_{Y_j}\|_1, \quad (6)$$

where $\|\cdot\|_1$ denotes the 1-norm of vector. Given the matching measures $\gamma(\mathbf{x}_i, \mathbf{y}_j)$ between all pairs of contour points, we initially construct a mapping $\bar{f} : \cup \mathcal{C} \rightarrow \cup \mathcal{S}$ by the following set of ordered pairs:

$$\bar{f} = \{\langle \mathbf{x}_i, \mathbf{y}_j \rangle \in \cup \mathcal{C} \times \cup \mathcal{S} : \gamma(\mathbf{x}_i, \mathbf{y}_j) < \epsilon_a\}, \quad (7)$$

where $\epsilon_a > 0$ is a constant representing the tolerant threshold of measure costs. This initial mapping \bar{f} is an augmented mapping, which may be a many-to-many correspondence. An ideal correspondence should be one-to-one, so we require further steps to reject outliers.

3.2 Refined Matching Using Global Constraints

For point-based matching of wide-baseline stereo, the epipolar is a mainly used global constraint. For contour-based matching, however, more global constraints are available. This is one of the reasons why contour-based stereo could outperform the point-based. We shall use the constraints of grouping, ordering and consistency to cooperate with the epipolar for refined matching.

The grouping constraint is important, since a contour is naturally a group of points. We state the grouping constraint as: Corresponding points should belong to corresponding contours (groups). The initially matched points in Section 3.1 may belong to unrelated contours; we expect that a pair of corresponding contours would have larger number of matched points than a pair of unrelated contours. Thus we assume that a pair of contours owning larger number of initially matched points has a higher probability to be corresponding contours, and the point matchings between them are more likely to be inlier matchings, see Fig. 2 for an illustration of this constraint.

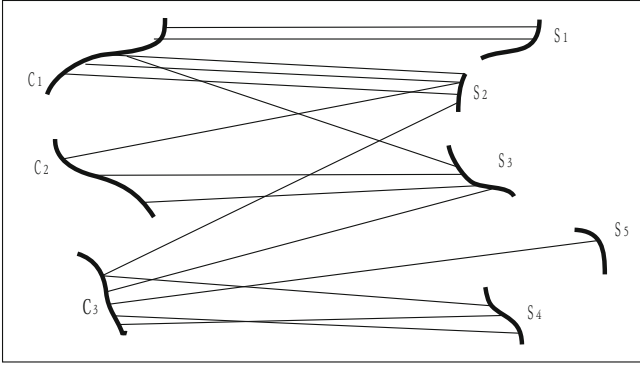


Fig. 2. Illustrating global constraints for contour matching. This figure simulates the initial matching result of Section 3.1. The grouping constraint indicates that the contour C_1 is more probably to match S_1 and S_2 , but less probably to match S_3 ; C_2 is more likely to match S_3 , and C_3 to S_4 . The ordering constraint indicates that the matching points between C_3 and S_4 contain outliers, since their matching orders are different.

We now formulate the grouping constraint in a simple way, so as to determine a mapping $f' : \cup C \rightarrow \cup S$ from the initial mapping \bar{f} as

$$f' = \{ \langle \mathbf{x}_i, \mathbf{y}_j \rangle \in \bar{f} : \langle \mathbf{x}_i, \mathbf{y}_j \rangle \in C_\alpha \times S_\beta, \text{card}(C_\alpha \times S_\beta) > \epsilon_g \}, \quad (8)$$

where $\text{card}(C_\alpha \times S_\beta)$ denotes the cardinal number of the set $C_\alpha \times S_\beta$, i.e., the number of matching pairs $\langle \mathbf{x}_i, \mathbf{y}_j \rangle$ between the contours C_α and S_β . The constant ϵ_g gives a threshold for the enough number of point matchings a pair of contours should own to become the corresponding contours. The mapping f' is thus resulted from rejecting outliers which do not satisfy the grouping constraint.

The ordering constraint means that matched points on contours should be ordered. If there is a matched point that does not follow the order of most other points, it has a higher probability to be an outlier. Fig. 2 also illustrates the ordering constraint. In experiments we find that this ordering constraint along one contour is less important than the grouping constraint among multiple contours, so we omit more details to formulate this constraint here.

The consistency constraint states that local transformations of corresponding contour points should be consistent with the global transformation of the entire image. We shall use the local transformations computed from the n points of local configurations to vote for a global transformation. A global transformation between stereo views is typically a perspective. As an intermediate step of rejecting outliers, however, we only use the local transformations to vote a global similarity transformation, and reject matching pairs whose local transformations do not consistent with the voted transform. This process is similar to the Hough transform commonly used in feature point matching.

We finally use the epipolar constraint. Since we have removed many outliers, the total number of matched points and the number of outliers may both be small relative to the initial matches. We thus use the deterministic parameter

estimation approach of constrained fundamental numerical scheme (CFNS) [4] to estimate the fundamental matrix between stereo views and reject outliers; this is slightly different from the widely used RANSAC-like schemes. More details of CFNS are given in [4] and we omit them for conciseness.

In the finally obtained correspondence mapping $f : \cup \mathcal{C} \rightarrow \cup \mathcal{S}$, if most of the matched points locate on silhouettes of objects between different views, and the silhouettes are not physically correspond to each other due to the viewpoint change, then the outlier rejection process may fail to remove these matches. This is a limitation of the contour-based approach; we shall illustrate it using experiments in the next section.

It has to explain that because of the length limit of the article, detailed mathematical explanation of the new constraints such as "ordering" and "consistency" is not given in this paper and will be presented in an extended work. The parameters used in the constraints and in the descriptor (such as the point number $n = 31$) are empirical and applicable to many experiments.

4 Experiments

We perform experiments in this section to demonstrate the performance of the proposed method and compare it with existing algorithms. The image data used in the experiments are divided into two parts: most are real-world images shot by us and others are standard test data in the community. The algorithm is implemented using Matlab. Contours should be extracted before using the proposed algorithm, but the contour detection algorithm does not be restricted. One can use the state-of-the-art edge detectors, such as those from Berkeley [13] or from Donoser et al. [5]. We generally use the latter due to the tradeoff between accuracy and speed. Contours can also be extracted from (closed) region boundaries by image segmentation algorithms, such as the graph-based segmentation [7] and the saliency driven segmentation [6]. The detected contours are typically interpolated using cubic spline to make them smoother for the matching purpose.

We first test the matching of interior contours of objects between stereo pairs. Interior contours refer to the contours formed by textures or structures on object surfaces, which are physically correspond between different views. The left part of Fig. 3 shows that the stereo correspondence can be well established using our algorithm to match interior contours of objects. The method of [5] is used to detect contours, but some of them are failed to be detected. This is a major reason why the number of matched contour points is relatively small. Next we test the matching of exterior contours of thin objects. Exterior contours (silhouettes) may be physically unrelated due to the change of view angles, but this effect is minor for thin objects. The right part of Fig. 3 depicts that our scheme matches well the points on silhouettes of thin objects, and the stereo correspondence is less affected by the bias of these silhouettes.

If objects are large enough, their silhouettes might have significant change between views. But using our affine invariant descriptor we cannot distinguish between interior and exterior contours; both of them could be matched, as illustrated in Fig. 4. In this test the images of carving have obvious silhouettes

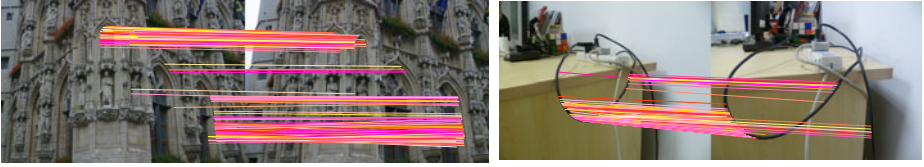


Fig. 3. Illustrating the matching of interior contours of objects (the left pair, images are from [18]) and exterior contours of thin objects (the right pair). Matched contour points are linked by the colored solid lines.

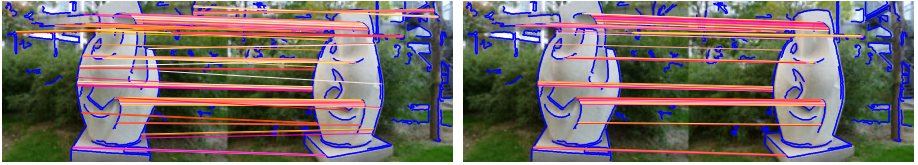


Fig. 4. Illustrating the matching of exterior contours (silhouettes). The blue points in images are detected contour points. The left pair is the matching result without the epipolar constraint. The right pair is the result after using the epipolar constraint; some matched points on unrelated silhouettes are maintained as inliers.

and many points on them are matched. After using epipolar constraint, matched points on the silhouettes cannot be fully removed; this may affect the estimated fundamental matrix. If more physically unrelated points on silhouettes are selected as inlier matchings, a wrong stereo estimation would be deduced. This is a major limitation of the contour-based stereo matching method and we shall improve it in further work.

The test images in the above figures are relatively simple. We now test real-world images with more clutters and larger depth of field, as shown in Fig. 5. We also compare our result with the SIFT-based stereo matching followed by an outlier rejection using the RANSAC algorithm. The results show that the SIFT-based matching is stable, but the number of detected and matched points are very small for images with less textures, e.g., the stereo pair of cars in the second row of Fig. 5. Our algorithm obtains larger number of matches as long as there are enough number of detected contours; this results in more stable and accurate estimation of fundamental matrices. Compared with point-based stereo matching like the SIFT, another advantage of our method is that it is able to match contour points of foreground objects within cluttered backgrounds. In the bottom two rows of Fig. 5, the SIFT-based matching cannot distinguish between foreground trees and their backgrounds, so it is difficult to recover the depth disparity between them; while the proposed method accurately matches contour points of both foreground and background, thus suitable for stereo views with large depth of field.

The computational load of the entire pipeline of the algorithm is comparable to the SIFT-based matching, but the dominant part of computational resources



Fig. 5. Experimental results for real-world images. All the left pairs are the results of SIFT-based stereo matching. All the right pairs are the results of the proposed algorithm. Matched contour points are linked by the colored solid lines. The blue contours of the top two and the bottom two rows of images are extracted using the method of [5] and [6] respectively.

Table 1. Comparison of computational time (seconds)

Test sets	Contour	Descriptor	Matching	Total	SIFT
Set 1	88.10	0.38	6.97	95.45	19.34
Set 2	10.96	0.25	5.28	16.49	26.27

is consumed by contour detection. We provide the test data of running times in Table 1, where the time data of Set 1 is the average of experiments in the top two rows of Fig. 5 and Set 2 is the average of experiments in the bottom two rows of that figure. The times for contour detection, descriptor construction and contour matching are listed in the first three columns; the total running times of our algorithm and the SIFT-based algorithm are listed in the last two columns. We expect that our algorithm can be improved if faster contour detection methods are applied. For example, the edge detector running on GPU [3] would be helpful to speed up the contour-based stereo matching.

In order to demonstrate more performance of our method, we test some stereo images from the Middlebury data sets [1], as shown in Fig. 6. The data sets are

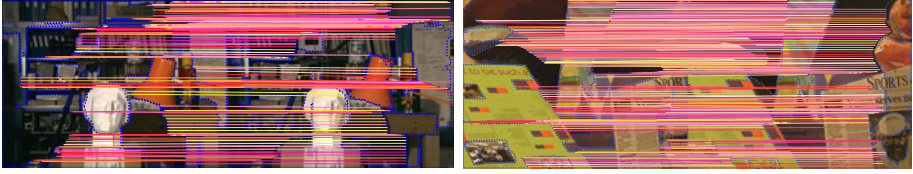


Fig. 6. Experimental results for stereo images in [1]

mainly used to evaluate short-baseline dense stereo matching algorithms. Though our contour-based algorithm only outputs sparse correspondence, it can be used as an initial step toward some dense correspondence, especially in the case of matching texture-less regions.

5 Conclusion

The proposed approach for wide-baseline stereo correspondence is based on the affine invariant description of local contour configurations, which approximates perspective transformations in stereo imaging. This contour-based method is superior to point-based matching such as the SIFT because it can obtain larger number of matches even for textureless objects, and accurately match contours of foreground objects within cluttered backgrounds, so as to recover their diverse depths. More global constraints for contour matching than the epipolar are also proposed, including the grouping, ordering and consistency constraints. They perform well for rejecting matching outliers.

A limitation of the proposed method is that it may generate matched points on physically unrelated silhouettes of objects, which can disturb the stereo estimation. The quality and speed of contour detectors affect the contour matching results; but the proposed is a flexible framework which can cooperate with other advanced contour detection methods. Improvements of the algorithm and detailed comparisons with more existing methods will be done in future work.

Acknowledgment

This work was supported by the National Natural Science Foundation of China under Grant 60803071.

References

1. Stereo data sets of middlebury college, <http://cat.middlebury.edu/stereo/data.html>
2. Begelfor, E., Werman, M.: Affine invariance revisited. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. II: 2087–2094 (2006)
3. Catanzaro, B., Su, B., Sundaram, N., Lee, Y., Murphy, M., Keutzer, K.: Efficient, high-quality image contour detection. In: International Conference on Computer Vision, pp. 2381–2388 (2009)

4. Chojnacki, W., Brooks, M., van den Hengel, A., Gawley, D.: On the fitting of surfaces to data with covariances. *IEEE Trans. Pattern Analysis and Machine Intelligence* 22(11), 1294–1303 (2000)
5. Donoser, M., Riemenschneider, H., Bischof, H.: Linked edges as stable region boundaries. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1665–1672 (2010)
6. Donoser, M., Urschler, M., Hirzer, M., Bischof, H.: Saliency driven total variation segmentation. In: *International Conference on Computer Vision*, pp. 817–824 (2009)
7. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. *International Journal of Computer Vision* 59(2), 167–181 (2004)
8. Golub, G.H., van Loan, C.F.: *Matrix Computations*, 3rd edn. Johns Hopkins University Press, Baltimore (1996)
9. Guan, S., Klette, R.: Belief-propagation on edge images for stereo analysis of image sequences. In: Sommer, G., Klette, R. (eds.) *RobVis 2008*. LNCS, vol. 4931, pp. 291–302. Springer, Heidelberg (2008)
10. Li, G., Zucker, S.W.: A differential geometrical model for contour-based stereo correspondence. In: *International Conference on Computer Vision* (2003)
11. Li, G., Zucker, S.W.: Contextual inference in contour-based stereo correspondence. *International Journal of Computer Vision* 69(1), 59–75 (2006)
12. Lowe, D.: Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision* 60(2), 91–110 (2004)
13. Maire, M., Arbelaez, P., Fowlkes, C., Malik, J.: Using contours to detect and localize junctions in natural images. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8 (2008)
14. Meltzer, J., Soatto, S.: Edge descriptors for robust wide-baseline correspondence. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. 1–8 (2008)
15. Mikolajczyk, K., Schmid, C.: Scale and affine invariant interest point detectors. *International Journal of Computer Vision* 60(1), 63–86 (2004)
16. Moallem, P., Faez, K., Haddadnia, J.: Reduction of the search space region in the edge based stereo correspondence. In: *International Conference on Image Processing*, pp. II: 149–152 (2001)
17. Srivastava, S., Ha, S., Lee, S., Cho, N., Lee, S.: Stereo matching using hierarchical belief propagation along ambiguity gradient. In: *International Conference on Image Processing*, pp. 2085–2088 (2009)
18. Strecha, C., Fransens, R., Gool, L.V.: Wide-baseline stereo from multiple views: a probabilistic account. In: *IEEE Conference on Computer Vision and Pattern Recognition*, pp. I: 552–559 (2004)
19. Tuytelaars, T., Van Gool, L.: Matching widely separated views based on affine invariant regions. *International Journal of Computer Vision* 59(1), 61–85 (2004)
20. Vedaldi, A., Soatto, S.: Features for recognition: viewpoint invariance for non-planar scenes. In: *International Conference on Computer Vision*, vol. 2, pp. 1474–1481 (2005)
21. Zuliani, M., Bhagavathy, S., Manjunath, B., Kenney, C.: Affine-invariant curve matching. In: *International Conference on Image Processing*, pp. V: 3041–3044 (2004)

Measuring the Coverage of Interest Point Detectors

Shoab Ehsan, Nadia Kanwal, Adrian F. Clark, and Klaus D. McDonald-Maier

School of Computer Science & Electronic Engineering,
University of Essex, Colchester CO4 3SQ UK
{sehsan,nkanwa,alien,kdm}@essex.ac.uk

Abstract. Repeatability is widely used as an indicator of the performance of an image feature detector but, although useful, it does not convey all the information that is required to describe performance. This paper explores the spatial distribution of interest points as an alternative indicator of performance, presenting a metric that is shown to concur with visual assessments. This metric is then extended to provide a measure of complementarity for pairs of detectors. Several state-of-the-art detectors are assessed, both individually and in combination. It is found that Scale Invariant Feature Operator (SFOP) is dominant, both when used alone and in combination with other detectors.

Keywords: Feature extraction, coverage, performance measure.

1 Introduction

The last decade has seen significant interest in the development of low-level vision techniques that are able to detect, describe and match image features [1,2,3,4,5,6]. The most popular of these algorithms operate in a way that makes them reasonably independent of geometric and photometric changes between the images being matched. Indubitably, the Scale Invariant Feature Transform (SIFT) [1] has been the operator of choice since its inception and has provided the impetus for the development of other techniques such as Speeded-Up Robust Features (SURF) [2] and Scale Invariant Feature Operator (SFOP) [6].

One of the main driving factors in this area is the improvement of detector performance. Repeatability [7,8], the ability of a detector to identify the same image features in a sequence of images, is considered a key indicator of detector performance and is the most frequently-employed measure in the literature for evaluating the performance of feature detectors [5,8]. However, it has been emphasized that repeatability is not the only characteristic that guarantees performance in a particular vision application [5,9]; other attributes, such as efficiency and the density of detected features, are also important. It is desirable to be able to characterize the performance of a feature detector in several complementary ways rather than relying only on repeatability [5,10,11].

One property that is crucial for the success of any feature detector is the spatial distribution of detected features, known as the *coverage* [10]. Many vision applications, such as tracking and narrow-baseline stereo, require a reasonably even distribution of detected interest points across an image to yield accurate results.

However, it is sometimes found that the features identified by detectors are concentrated on a prominent textured object, a small region of the image. Robustness to occlusion, accurate multi-view geometry estimation, accurate scene interpretation and better performance on blurred images are some of the important advantages of detectors whose features cover images well [10,11].

Despite its significance, there is no standard metric for measuring the coverage of feature detectors [10]. An approach based on the convex hull is employed in [12] to measure the spatial distribution for evaluating feature detectors. However, a convex hull traces the boundary of interest points without considering their density, resulting in an over-estimation of coverage. In [13], a completeness measure is presented but requires more investigation due to its dependence upon the entropy coding scheme and Gaussian image model used, and may provide varying results with other coding schemes for different feature types.

To fill this void, this paper presents a metric for measuring the spatial distribution of detector responses. It will be shown that the proposed measure is a reliable method for evaluating the performance of feature detectors. Since complementary feature detectors (*i.e.*, combining detectors that identify different types of feature) are becoming more popular for vision tasks [14,15,16], it is important to have measures of complementarity for multiple feature detectors, so that their combined performance can be predicted and measured [5]. This paper shows how *mutual coverage*, the coverage of a combination of interest points from multiple detectors, can be used to measure complementarity.

The rest of the paper is structured as follows: Section 2 describes the coverage measure, which is used to evaluate the performances of eleven state-of-the-art detectors on well-established data sets in Section 3. A complementarity measure derived from coverage, mutual coverage, is proposed in Section 4 and its effectiveness is demonstrated by results for combination of detectors. Finally, conclusions are presented in Section 5.

2 Measuring Coverage

There are several *desiderata* for a coverage measure:

- differences in coverage should be consistent with performance differences obtained by visual inspection;
- penalization of techniques that concentrate interest points in a small region; and
- avoidance of overestimation by taking into account the density of feature points.

The obvious way to estimate coverage is to calculate the mean Euclidean distance between feature points. However, different densities of feature points yield the same mean Euclidean distance. Conversely, the harmonic mean, which is widely used in data clustering algorithms [17], does penalize closely-spaced feature points, which augurs well for encapsulating their spatial distribution. Indeed, the harmonic mean is an inherently conservative approach for estimating the central tendency of a sample space, as:

$$A(x_1, \dots, x_n) \geq G(x_1, \dots, x_n) \geq H(x_1, \dots, x_n) \quad (1)$$

where $A(\cdot)$ is the arithmetic, $G(\cdot)$ the geometric and $H(\cdot)$ the harmonic mean of the sample set $x_1, \dots, x_n, x_i \geq 0 \forall i$.

Formally, we assume that p_1, \dots, p_N are the N interest points detected by a feature detector in image $I(x, y)$, where x and y are the spatial coordinates. Taking p_i as a reference interest point, the Euclidean distance d_{ij} between p_i and some other interest point p_j is

$$d_{ij} = \sqrt{(x_i - x_j)^2 + (y_i - y_j)^2} \quad (2)$$

providing $i \neq j$. Computation of (2) provides $N - 1$ Euclidean distances for each reference interest point p_i . The harmonic mean of d_{ij} is then calculated to obtain a mean distance $D_i, i = 1, \dots, N$ with p_i as reference:

$$D_i = \frac{N - 1}{\sum_{j=1, j \neq i}^N \left(\frac{1}{d_{ij}} \right)} \quad (3)$$

Since the choice of the reference interest point can affect the calculated Euclidean distance, this process is repeated using each interest point as reference in turn, resulting in a set of distances D_i . Finally, the coverage of the feature detector is calculated as

$$\frac{N}{\sum_{i=1}^N \left(\frac{1}{D_i} \right)} \quad (4)$$

Since multi-scale feature detectors may provide image features at exactly the same physical location but different scales, interest points that result in zero Euclidean distance in (2) are excluded from these calculations on the basis that they do not provide independent evidence of an interest point.

In general, a large coverage value is desirable for a feature detector as a small value implies the concentration of interest points into a small region. However, the final coverage value obtained from (4) needs to be considered against the dimensions of a specific image as the same coverage value may indicate good distribution for a small image but poor distribution for a large one.

3 Performance Evaluation

For the proposed coverage measure to have any value, its values need to be consistent with visual assessments of coverage across a range of feature detectors and a variety of images. To that end, this section presents a comparison of the coverage of eleven state-of-the-art feature detectors: SIFT (Difference-of-Gaussians), SURF (Fast Hessian), Harris-Laplace, Hessian-Laplace, Harris-Affine, Hessian-Affine, Edge-based Regions (EBR), Intensity-based Regions (IBR), Salient Regions, Maximally Stable Extremal Regions (MSER) and Scale Invariant Feature Operator (SFOP) [5,6].

Although different parameters of a feature detector can be varied to yield more interest points, it has a negative effect on repeatability and performance [13]. Therefore, authors’ original binaries have been utilized, with parameters set to values recommended by them, and the results presented were obtained with the widely-used Oxford datasets [18]. The parameter settings and the datasets used make our results a direct complement to existing evaluations.

To demonstrate the effectiveness of this coverage measure, first consider the case of Leuven dataset [18] in Fig. 1. It is evident that SFOP outperforms the other detectors, where as values for EBR, Harris-Laplace and Harris-Affine indicate a poor spatial distribution of interest points. To back up these results, the actual distribution of detector responses for SFOP, IBR, Harris-Laplace and EBR for image 1 of the Leuven dataset are presented in Fig. 2. Visual inspection of these distributions is consistent with the coverage results of Fig. 1.

The coverage values obtained for Boat dataset [18] are presented in Fig. 3. Again, the performance of well-established techniques like SIFT and SURF is eclipsed by SFOP, a relatively new entrant in this domain. Other popular methods, such as Harris-Laplace, Harris-Affine, Hessian-Affine and EBR, again fare poorly. In addition, the curves depicted in Fig. 1 and 3 also exemplify the effects of illumination changes (Leuven) and zoom and rotation (Boat) on coverage.

A summary of the mean results obtained with all these feature detectors for the remaining datasets [18] is presented in Table 1. It is clear that SFOP achieves much better coverage than the other feature detectors for almost all datasets under various geometric and photometric transformations.

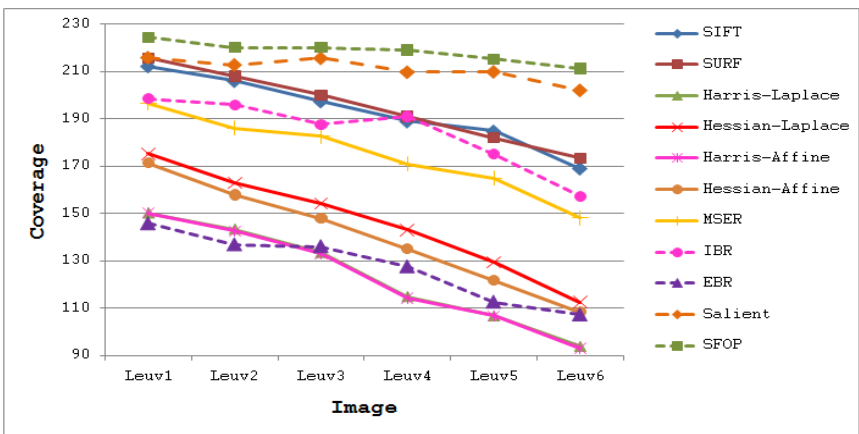


Fig. 1. Coverage results for Leuven dataset [18]

To exemplify the impact of these results on real-world applications, consider the task of homography estimation for the Leuven dataset. The mean error was computed between the positions of points projected from one image to the other, using a

‘ground-truth’ homography from [18], and a homography determined using the above detectors. SFOP performed the best, with a mean error of 0.245, where as EBR achieved a poor value of 3.672, consistent with the results shown in Fig. 1 and 2. In addition, we refer the reader to [11] that explains the significance of coverage of interest points (including those that cannot be matched accurately) for the task of scene interpretation. The proposed measure seems a viable method for determining coverage for such applications.

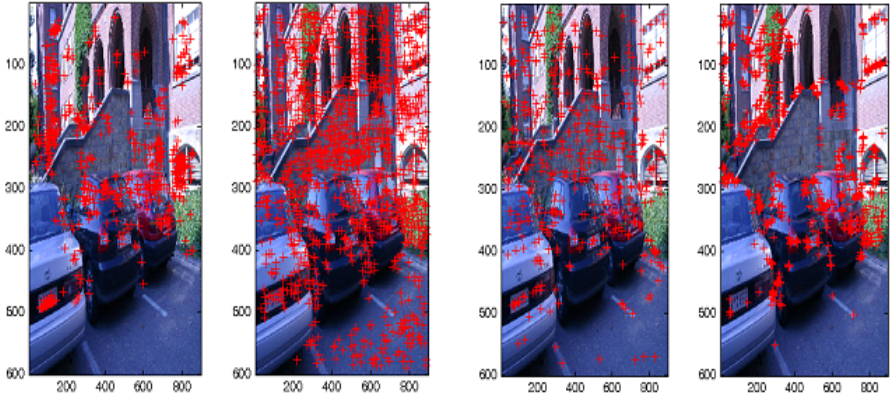


Fig. 2. Actual detector responses for image 1 of Leuven dataset [18]. From left to right: EBR, SFOP, IBR and Harris-Laplace.

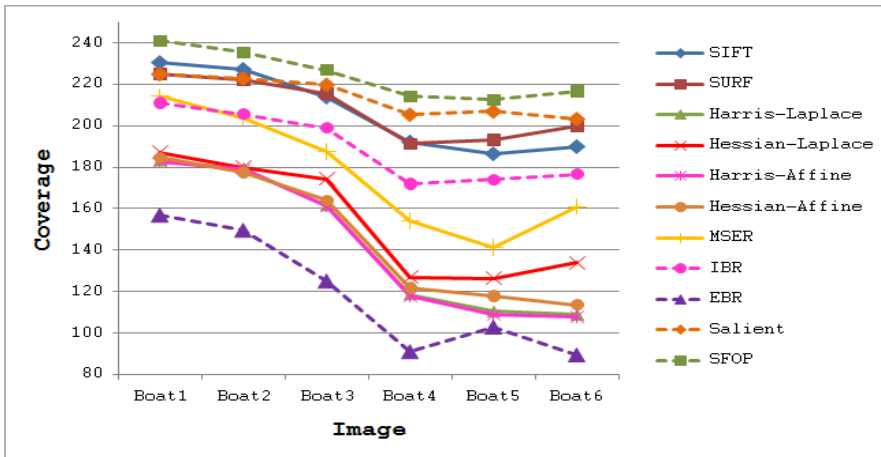


Fig. 3. Coverage results for the Boat dataset [18]

Table 1. Coverage results for state-of-the-art feature detectors

	Bark	Bikes	Graffiti	Trees	UBC	Wall
SIFT(DoG)	190.3	207.8	221.0	263.4	204.2	253.5
SURF(FH)	195.8	228.1	221.9	265.4	205.4	246.6
Harris-Lap	122.9	136.5	181.2	230.2	154.5	213.7
Hessian-Lap	120.0	154.5	199.2	234.2	154.9	208.6
Harris-Aff	122.8	136.0	181.0	229.9	153.8	212.8
Hessian-Aff	119.9	148.9	191.0	233.0	153.5	208.2
Salient Regions	190.6	258.7	218.0	256.4	201.5	236.4
EBR	139.2	138.3	166.4	214.3	119.0	204.4
IBR	192.3	214.7	209.7	255.5	198.4	243.8
MSER	179.6	86.4	200.3	229.6	200.6	248.3
SFOP	204.4	246.3	228.7	270.3	213.8	256.5

4 Mutual Coverage for Measuring Complementarity

Since the utilization of combinations of feature detectors is an emerging trend in local feature detection [5], this section proposes a new measure based on coverage to estimate how well these detectors complement one another. In addition to the principles mentioned in Section 2, the objective here is to penalize techniques that detect several interest points in a small region of an image. If detector A and detector B detect most feature points at same physical locations, they should have a low complementarity score. Conversely, a high score should be achieved if detector A and detector B detect most features at widely-spaced physical locations, indicating that they complement each other well. Again, a metric utilizing the harmonic mean seems a promising solution to achieve the required goal.

Formally, let us consider an image $I(x, y)$, where x and y are the spatial coordinates, being operated on by M feature detectors F_1, F_2, \dots, F_M , so that $P_z = \{P_{z1}, P_{z2}, \dots, P_{zN}\}$ is the set of N feature points detected by F_z . We then define

$$P_{zk} = P_z \cup P_k \quad (5)$$

as the set of feature points detected in image $I(x, y)$ by F_z and F_k . The coverage is then calculated as described in Section 2 using P_{zk} ; as that includes points detected by both F_z and F_k , we denote it as the *mutual coverage* of F_z and F_k for image $I(x, y)$. Although this paper confines itself to combinations of two detectors, this notion of mutual coverage can be extended to more than two by simply combining their feature points in (5).

Mutual coverage has been applied to combinations of the detectors examined in the previous section. Inspired by [13], they can be categorized into four major classes, shown in Table 2. For the purpose of this work, we confine ourselves to combinations of two detectors selected from two different categories; for example, SIFT is combined with EBR but not with SURF as they both detect blobs in a given image.

Fig. 4, 5 and 6 depict the average image coverage for SFOP, EBR and MSER when grouped with detectors from other categories for all 48 images of the Oxford datasets

[18]. Interestingly, these results are consistent with the completeness results presented in [13]. Detectors from other categories perform well when combined with SFOP. The best results are achieved by grouping SFOP with a segmentation-based detector. A corner detector combined with a blob detector (except Hessian-Laplace and Hessian-Affine) yields good coverage. Segmentation-based detectors, however, do not seem to work well with corner detectors.

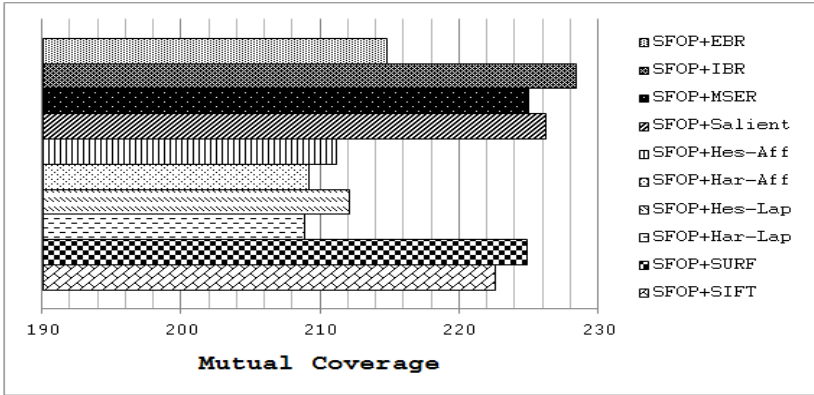


Fig. 4. Mutual coverage of SFOP in combination with other detectors

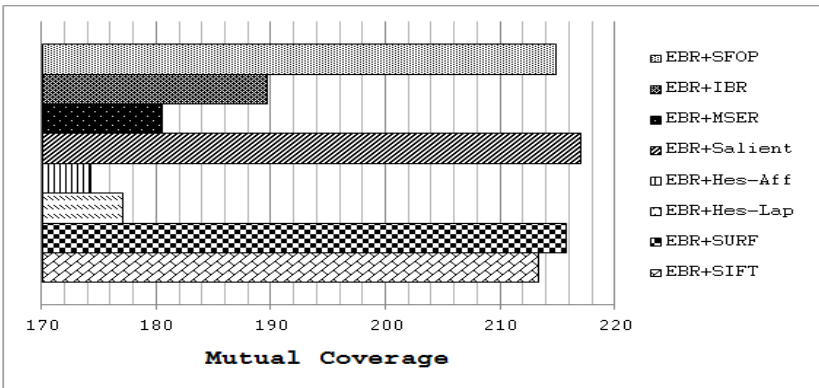


Fig. 5. Mutual coverage of EBR in combination with other detectors

Table 2. A taxonomy of state-of-the-art feature detectors

Category	Type	Detectors
1.	Blob detectors	SIFT, SURF, Hessian-Laplace, Hessian-Affine, Salient Regions
2.	Spiral detectors	Scale Invariant Feature Operator
3.	Corner detectors	EBR, Harris-Laplace, Harris-Affine
4.	Segmentation-based detectors	MSER, Intensity-based Regions

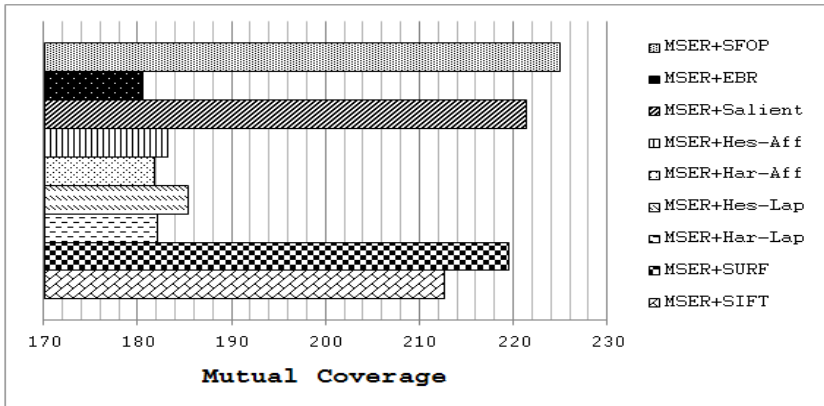


Fig. 6. Mutual coverage of MSER in combination with other detectors

5 Conclusions

The performance of any image feature detector is dependent upon a number of different characteristics and one such property is coverage. This paper has proposed a coverage measure that produces results consistent with visual inspection. Furthermore, the mutual coverage of several feature detectors can be obtained simply by concatenating the feature points they detect and calculating the coverage of the combination. This gives us a rapid, principled way of determining whether combinations of interest point detectors will be complementary without having to undertake extensive evaluation studies; indeed, calculation is so rapid that one can consider using it online in an intelligent detector that adds features from other detectors in order to ensure that coverage, and hence accuracy of subsequent processing, is good enough.

An examination of the coverages of a range of state-of-the-art detectors identifies SFOP as the outstanding detector, both individually and when used in combination with other detectors.

Acknowledgment. This work was supported in part by the UK EPSRC under grant EP/I500952/1.

References

1. Lowe, D.: Distinctive Image Features from Scale-Invariant Keypoints. *International Journal of Computer Vision* 60, 91–110 (2004)
2. Bay, H., Ess, A., Tuytelaars, T., Van Gool, L.: Speeded-Up Robust Features (SURF). *Computer Vision and Image Understanding* 110, 346–359 (2008)
3. Matas, J., Chum, O., Urban, M., Pajdla, T.: Robust Wide Baseline Stereo from Maximally Stable Extremal Regions. In: *BMVC*, Cardiff, UK, pp. 384–393 (2002)
4. Kadir, T., Zisserman, A., Brady, M.: An affine invariant salient region detector. In: Pajdla, T., Matas, J.(G.) (eds.) *ECCV 2004*. LNCS, vol. 3021, pp. 228–241. Springer, Heidelberg (2004)

5. Tuytelaars, T., Mikolajczyk, K.: Local Invariant Feature Detectors: A Survey. *Foundations and Trends in Computer Graphics and Vision* 3, 177–280 (2007)
6. Forstner, W., Dickscheid, T., Schindler, F.: Detecting Interpretable and Accurate Scale-Invariant Keypoints. In: ICCV, Kyoto, Japan, pp. 2256–2263 (2009)
7. Schmid, C., Mohr, R., Bauckhage, C.: Evaluation of Interest Point Detectors. *International Journal of Computer Vision* 37, 151–172 (2000)
8. Ehsan, S., Kanwal, N., Clark, A., McDonald-Maier, K.: Improved Repeatability Measures for Evaluating Performance of Feature Detectors. *Electronics Letters* 46, 998–1000 (2010)
9. Nowak, E., Jurie, F., Triggs, B.: Sampling Strategies for Bag-of-Features Image Classification. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3954, pp. 490–503. Springer, Heidelberg (2006)
10. Perdoch, M., Matas, J., Obdrzalek, S.: Stable Affine Frames on Isophotes. In: ICCV, Rio de Janeiro, Brazil (2007)
11. Tuytelaars, T.: Dense Interest Points. In: CVPR, San Francisco, USA, pp. 2281–2288 (2010)
12. Dickscheid, T., Förstner, W.: Evaluating the Suitability of Feature Detectors for Automatic Image Orientation Systems. In: Fritz, M., Schiele, B., Piater, J.H. (eds.) ICVS 2009. LNCS, vol. 5815, pp. 305–314. Springer, Heidelberg (2009)
13. Dickscheid, T., Schindler, F., Förstner, W.: Coding Images with Local Features. *International Journal of Computer Vision* (2010), doi: 10.1007/s11263-010-0340-z
14. Lazebnik, S., Schmid, C., Ponce, J.: Sparse Texture Representation using Affine-Invariant Neighborhoods. In: IEEE CVPR, Wisconsin, USA, pp. 319–324 (June 2003)
15. Mikolajczyk, K., Leibe, B., Schiele, B.: Multiple Object Class Detection with a Generative Model. In: IEEE CVPR, New York, USA, pp. 26–36 (2006)
16. Sivic, J., Zisserman, A.: Video Google: A Text Retrieval Approach to Object Matching in Videos. In: ICCV, Nice, France, vol. 2, pp. 1470–1477 (2003)
17. Zhang, B., Hsu, M., Dayal, U.: K-Harmonic Means-A Spatial Clustering Algorithm with Boosting. In: Roddick, J., Hornsby, K.S. (eds.) TSDM 2000. LNCS (LNAI), vol. 2007, pp. 31–45. Springer, Heidelberg (2001)
18. Oxford Data Sets, <http://www.robots.ox.ac.uk/~vgg/research/affine/>

Non-uniform Mesh Warping for Content-Aware Image Retargeting

Huiyun Bao and Xueqing Li

School of Computer Science and Technology, Shandong University, Jinan,
Shandong, 250101, China
baohuiyun@gmail.com, xqli@sdu.edu.cn

Abstract. Image retargeting is the process of adapting an existing image to display with arbitrary sizes and aspect ratios. A compelling retargeting method aims at preserving the viewers' experience by maintaining the significant regions in the image. In this paper, we present a novel image retargeting method based on non-uniform mesh warping, which can effectively preserve both the significant regions and the global configuration of the image. The main idea of our method is sampling mesh vertices based on the saliency map, that is to say, we place mesh vertices more densely in the significant regions, defining different quadratic error metrics to measure image distortion and adopting a patch-linking scheme that can better preserve the global visual effect of the entire image. Moreover, to increase efficiency, we formulate the image retargeting as a quadratic minimization problem carried out by solving linear systems. Our experimental results verify its effectiveness.

Keywords: Image retargeting, sampling mesh vertices, non-uniform mesh warping, patch-linking scheme.

1 Introduction

With the proliferation of display devices, such as television, notebooks, PDAs and cell phones, adjusting an image to heterogeneous devices with different sizes and aspect ratios is becoming more attractive. The critical problem for image retargeting is how to retarget the image effectively and to prevent the prominent object of the image from distorting. To address this problem, a large amount of effort has been spent on image retargeting.

Previous approaches mainly include cropping and scaling images. Cropping is to crop the input image and get a cropping with the same aspect ratio as the target display. These methods inevitably discard too much information. Scaling can be performed in real-time and can preserve the global configuration. However, scaling image to arbitrary aspect ratios either stretch or squash the significant regions.

Seam carving [1, 3] is an efficient technique for content-aware image resizing, which works by greedily carving out or inserting one-dimensional seams passing through unimportant regions. The drawback of this method is that the global configuration of an image may be severely damaged due to the energy-based strategy

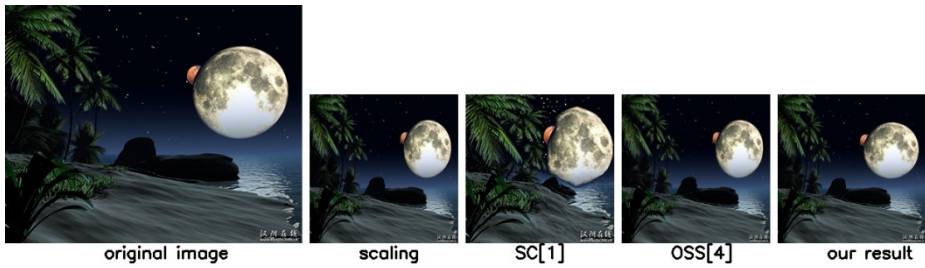


Fig. 1. An example of retargeting: input size is 400x320 and target size is 200x200. The results are computed using scaling, seam carving [1], optimal scale-and-stretch [4] and our method respectively.

of the algorithm that always removes the seams containing or removing low energy until the desired image size is achieved. Besides, because the method adopts dynamic programming for seam searching, the computational speed is low.

Wang et al. [4] provided an optimized scale-and-stretch warping method for image resizing using a quad-mesh, which attempts to ensure that important quads have homogeneous scaling while minimizing bending of grid line. This method distributes the distortion in all spatial directions. Compared to early image warping approaches, it better utilizes the available homogeneous regions to absorb the distortion. However, this method is lack of large scale feature preservation because a salient object occupies many quads but each quad has a locally acceptable homogeneous scaling.

In this paper, we propose a novel content-aware image retargeting method that automatic samples mesh vertices based on the saliency and designs quadratic error metrics over the mesh to measure different image distortion and adopt patch-linking scheme [15] which apply constraints to neighboring meshes with similar significance to link image patches together. By using our method, the distortion is better diffused and the global configuration is better preserved, as shown in Fig.1.

Firstly, we calculate the saliency map of an image combining the graph-based visual saliency and face information, and then sample mesh vertices based the saliency map, finally compute the error distortion metrics using this spatially varying importance. In this way, the retargeting problem is formulated as a quadratic minimization problem that can be solved using linear system. Finally, we render the final result using texture mapping [7].

Our main contributions are as follows:

- We propose a novel saliency-driven approach to automatic sample the non-uniform mesh.
- A patch distance measure between two neighbor mesh which is used when calculating the mesh-link scheme.

2 Related Work

Image retargeting is a standard tool in many image processing applications. Recently, many methods have appeared in the literature for retargeting images to displays with

different resolutions and aspect ratios. Traditional methods just work by uniformly resizing the image to a target size without taking the image content into account, equally propagating the distortion throughout the entire image and noticeably squeezing salient objects. To solve this problem, many approaches attempt to remove the unimportant information from the image periphery [8, 12]. Based on a face detection technique [14] and a saliency measure [9], the image is cropped to fit the target aspect ratio and then uniformly resized by traditional interpolation. More sophisticated cropping methods usually require human intervention to create an optimal window for the most appropriate portion of the scene. These methods work well for some special applications. However, cropping methods may potentially remove significant objects next to the image boundary, especially when the output resolution is lower than the input resolution.

Recently proposed retargeting methods try to preserve the prominent object while reducing or removing other image content. Seam carving methods [1, 3, 10] reduce or expand monotonic 1D seam of pixels that run roughly in the orthogonal direction in a certain direction. To reduce artifacts, they search for minimal-cost seams that pass through homogeneous regions by calculating their forward [1] or backward energy [3]. These methods can produce very impressive results, but may deform prominent object when the homogeneous information in the required spatial direction runs out, especially structural object. Moreover, the global configuration may also be damaged in the output.

Image warping [2, 4, 5, 6, 11, 15, 18] provides a continuous solution to image retargeting. To reduce the output distortion, the warping functions are generally acquired by a global optimization that squeezes or stretches homogeneous regions. Gal et al. [11] warp an image into various shapes, enforcing the user specified features to undergo similarity transformations that employ a simple heuristic to determine the scaling of the marked features. Wolf et al. [2] and Wang et al. [5] automatically determine the significance of each pixel and merge of the pixels of lesser importance in the reduction direction. Wang et al. [4] propose a “scale-and-stretch” warping method that is iteratively updates a warped image that matches optimal local scaling factors, but because the distortion is distributed in all spatial directions, some objects may be excessively distorted, damaged the global configuration of the original image. Zhang et al. [6] estimate a nonlinear warping by minimizing a quadratic distortion energy function defined over a set of control points, including the vertices of a regular mesh grid and a lot of selected edge points, and are grouped into small local groups called handles, which are warped using a linear similarity transformation. Wang et al. [13] calculated the mesh similarity invariance of local region based on triangle similarity, constructed a quadratic energy to measure the similarity error of each local region and finally obtained by minimizing the energy function sum with salience as weight, which can preserve the shape of the local prominent regions. Niu et al [15] defined a variety of quadratic metrics to measure image distortion, introduced a patch-linking scheme, designed different strategies for upsizing and downsizing and formulated image resizing as a quadratic minimization problem, this method performs impressively and can effectively preserve the shape of the important objects and the global configuration of the image.

3 Image Retargeting Using Non-uniform Mesh

We offer an image retargeting method that constructs the non-uniform mesh of the original image and calculates the target mesh position with minimum quadratic distortion by solving a linear optimization problem.

We calculate a significance map and spread the distortion according to significant value of each patch, like the previous method [4, 15]. The significance map is composed of edge energy, graph-based visual saliency [17] and face detection information [14]. To keep the global visual effect, we measure not only the distortion of each image mesh, but also that between neighboring image grid mesh that satisfy a certain condition. Quadratic metrics [15] are designed to measure image distortion in different types.

We first calculate the significance map of an image, then represent the image as a mesh $G = (V, E, F)$ with vertices V , edges E and quad faces F , where $V = [v_1^T, v_2^T, \dots, v_n^T]$, $v_i \in \mathbb{R}^2$ denotes the initial vertex coordinates. V and E form horizontal and vertical grid lines partitioning the image into F . We solve the problem of finding a new mesh $V' = [v_1'^T, v_2'^T, \dots, v_n'^T]$ with minimal distortion. The mesh is constructed based on the significance map, that is to say, the square of the mesh is less (more densely) in important region and is greater in unimportant region.

In the following subsections, firstly, we describe calculating significance map, secondly we describe how to construct the non-uniform mesh and then we give the measure metrics on different distortion of the mesh and solve the distortion energy minimization problem.

3.1 Significance Calculation

It is important to get accurate significance map of image for understanding the image content. We present a new method to compute the significance map. Firstly, we compute saliency map by adopting GBVS [17] which is a new bottom-up visual saliency model based a simple, biologically plausible, and distributed computation. Compared with Itti's [9], this method can differentiate the important region and prominent objects, but the boundary is fuzzy which may lead to damage the important regions, as shown in Fig.2. To enhance the boundary information, we introduce the edge energy into the significance map. We extract the edge information by using canny operator.

We also want to prevent the faces from distorting, because people are very sensitive to the distortion of faces. So we add face information to our significance map. Currently, we use the Viola and Jones face detection mechanism [14]. The detector returns a list of detected faces, and then we adapt the cubic function [2] to compute the weight of each detected face, as shown in Fig.2 (e).

Finally, we obtain the final significance map by combining the saliency map, edge energy and face information as follows:

$$S(x, y) = \max(S_s(x, y), S_e(x, y), S_f(x, y)) . \quad (1)$$

where $S(x, y)$ denotes the significance value that is at pixel (x, y) in the original image, $S_s(x, y)$ is the saliency value, $S_e(x, y)$ is the edge energy and $S_f(x, y)$ is the face saliency value.

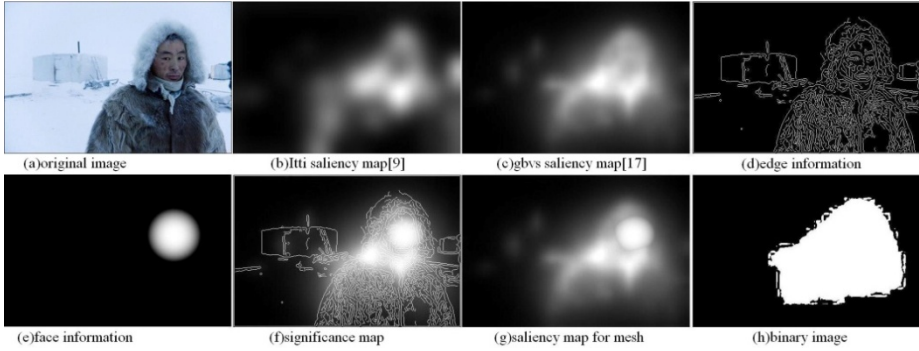


Fig. 2. An example of computing significance map: (a) original image;(b) Itti's[9] saliency map;(c)gbvs [17] saliency map;(d)edge information computing by canny operator;(e) face information detected by[16];(f) the final significance map;(g)the significance map for constructing the mesh;(h)the binary image of the significance map, $p=0.75$

3.2 Grid Mesh Construction

The finer is the mesh, the less distortion is the quad. We present a novel method to construct the mesh based on the significance value of each pixel. We first obtain the significance value for constructing the mesh by combining the saliency map $S_s(x, y)$ and the face information $S_f(x, y)$, and then binary the significance map image as shown in Fig.2(h). The function is as follows:

$$\begin{aligned}
 S_b(x, y) &= \begin{cases} 1, & S_g(x, y) \geq S_t; \\ 0, & \text{otherwise.} \end{cases} \\
 S_g(x, y) &= [\max(S_s(x, y), S_f(x, y))]^2 \\
 S_t &= \frac{1}{m * n} \sum_x \sum_y S_g(x, y) * p .
 \end{aligned} \tag{2}$$

where $S_b(x, y)$ denotes the binary value of each pixel, S_t is the threshold value, m and n is the size of the original image, and p is the percent factor to compute the threshold value. The greater is the p , the more coverage is the important region.

We construct the mesh based on the binary image. Firstly, we compute the mesh vertices in the y direction $V_y = [v_{y1}^T, v_{y2}^T, \dots, v_{yN}^T], N \in \mathbb{R}^2$, and then calculate the mesh vertices in the x direction $V_x = [v_{x1}^T, v_{x2}^T, \dots, v_{xM}^T], M \in \mathbb{R}^2$ on the basis of V_y . We compute V_y just in the first column. The procedure is as follow:

1. Set the size of the largest mesh GX_0 and GY_0 , and denote the largest patch as M .
2. Construct the position of the i_{th} vertice in the y direction, denoted as v_{yi} . Firstly, we compute the average value of the significance value inside the i_{th} mesh M , denoted as S_{Qi} , if S_{Qi} is less than the constraint value, presented as \mathcal{E} ($\mathcal{E}=0.1$), we set $v_{yi} = v_{y_{i-1}} + GY_0$, otherwise $v_y = v_{y_{i-1}} + GY_0 * C_y$.
3. Compute the position of the vertices similar to 2 until the position of the vertice is the height of the image.

After we get the V_y , we then compute the V_x similar to computing the V_y , but we will compute all the rows. The functions are as follows:

$$\begin{cases} S_{Q_i} = \frac{1}{GX_0 * GY_0} \sum_{x=1}^{GX} \sum_{y=y_{i-1}}^{(y_{i-1} + GX_0)} S_b(x, y) \\ v_{y_i} = \begin{cases} v_{y_{i-1}} + GY_0, & S_{Q_i} \leq \varepsilon \\ v_{y_{i-1}} + GY_0 * c_y, & \text{otherwise} \end{cases} \end{cases}, i \in [1, 2, \dots, N]. \quad (3)$$

$$\begin{cases} S_{Q(i,j)} = \frac{1}{GX_0 * (v_{y_{i+1}} - v_{y_i})} \sum_{x=v_x(j-1)}^{(v_x(j-1) + GX_0)} \sum_{y=v_{y_i}}^{v_{y_{i+1}}} S_b(x, y), \quad v_{y_i} \in V_y \\ v_{x_j} = \begin{cases} v_{x_{j-1}} + GX_0, & S_{Q(i,j)} \leq \varepsilon \\ v_{x_{j-1}} + GX_0 * c_x, & \text{otherwise} \end{cases} \end{cases}, j \in [1, 2, \dots, M] \quad (4)$$

where c_x and c_y are the ratio of sampling the vertices in the x and y direction respectively. $S_{Q(i,j)}$ presents the quad significance value of the quad in the i_{th} row and the j_{th} column.

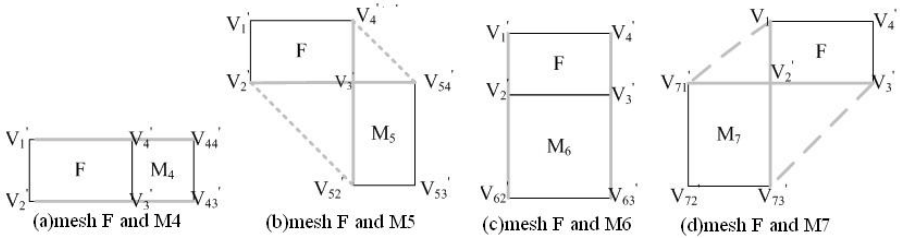


Fig. 3. The relationship between mesh F and its four neighboring meshes: (a) F and the right neighboring mesh M_4 , (b) F and the right bottom mesh M_5 , (c) F and the bottom mesh M_6 , (d) F and the left bottom mesh M_7

3.3 Distortion Metrics

Different quadratic metrics, including shape, orientation and scale distortion, are defined to measure different image distortion on the mesh. And then an optimization problem is solved to retarget image with minimum visual distortion. To prevent the shape from distorting, we make each image mesh to undergo only similarity transformation. Because we are sensitive to the orientation of the important content, we expect preserving the orientation with the minimum distortion. Moreover, we hope the important object to undergo a uniform scaling, and the scaling factor is determined based on the source and target image size.

Because an individual image mesh cannot keep the overall image configuration well [15], we adopt constraints to every two neighboring meshes with similar significance to link image meshes together, which can effectively keep the global configuration.

The objective function, measuring the shape, orientation and scale distortion, is discussed in [15], which computes all the patch-neighboring and works well, but it is at the expense of time. Therefore, we think that it is sufficient for keeping the entire important region to compute the two neighboring patches with similar energy, which can save a large amount of time.

3.3.1 Shape Distortion

The single shape distortion associated with mesh $F\{v_1, v_2, v_3, v_4\}$ is:

$$E_s\{v_1, v_2, v_3, v_4\} = \sum_{(i=1,2,3,4)} \|v_i^d - v_i\|^2 * S_m . \tag{5}$$

where S_m is the sum of the significance value inside the mesh $F\{v_1, v_2, v_3, v_4\}$

Niu et al. [15] encourage the eight neighboring patches to undergo the same transformation as itself for each mesh $F\{v_1, v_2, v_3, v_4\}$ in F . This process consumes a large amount of time, so we encourage the neighboring meshes with similar importance values to undergo the same transformation. The neighboring relationship between two patches is symmetric, for example, M_5 is a neighbor of $F\{v_1, v_2, v_3, v_4\}$ and $F\{v_1, v_2, v_3, v_4\}$ is also a neighbor of M_5 ; we only consider the four neighboring meshes of mesh $F\{v_1, v_2, v_3, v_4\}$, M_4, M_5, M_6 and M_7 , as shown in Fig.3. Firstly, we measure the similarity between the two neighboring patches, and then determine whether or not adopt the objective function terms to measure the shape distortion for every patch link.

$$E_{SL}\{v_1, v_2, v_3, v_4\} = \begin{cases} E_{SL}\{v_1, v_2, v_3, v_4\} + E_{SL4}, & |S_F - S_{M4}| \leq d \\ E_{SL}\{v_1, v_2, v_3, v_4\} + E_{SL5}, & |S_F - S_{M5}| \leq d \\ E_{SL}\{v_1, v_2, v_3, v_4\} + E_{SL6}, & |S_F - S_{M6}| \leq d \\ E_{SL}\{v_1, v_2, v_3, v_4\} + E_{SL7}, & |S_F - S_{M7}| \leq d \end{cases} \tag{6}$$

where $E_{SL}\{v_1, v_2, v_3, v_4\}$ presents the shape distortion of the patch link, d denotes the maximum similarity difference. We don't apply patch-linking scheme to the neighboring patches whose similarity difference are bigger than d . $S_F, S_{M4}, S_{M5}, S_{M6}$ and S_{M7} denote the significance value of the quad mesh F, M_4, M_5, M_6 and M_7 respectively. $E_{SL4}, E_{SL5}, E_{SL6}$ and E_{SL7} are calculated by using the function terms in [15].

3.3.2 Orientation Distortion

Orientation distortion is another distortion that disturbs the visual effect. The single orientation distortion associated with quad $F\{v_1, v_2, v_3, v_4\}$ is:

$$E_o\{v_1, v_2, v_3, v_4\} = (E_{OH} + E_{OV}) * S_m . \tag{7}$$

where S_m is the sum of the significance value inside the mesh $F\{v_1, v_2, v_3, v_4\}$, E_{OH} and E_{OV} are the difference of two end points of the horizontal lines at vertical direction and the difference of two end points of the vertical lines at horizontal direction respectively, adapting the function terms of [15].

To avoid orientation distortion, we also adopt orientation links. Similar to shape links distortion, we only consider the four neighboring meshes of mesh $F\{v_1, v_2, v_3, v_4\}$, M_3, M_4, M_5 and M_7 . The function is as follow:

$$E_{OL}\{v_1, v_2, v_3, v_4\} = \begin{cases} E_{OL}\{v_1, v_2, v_3, v_4\} + E_{OL4}, & |S_F - S_{M4}| \leq d \\ E_{OL}\{v_1, v_2, v_3, v_4\} + E_{OL5}, & |S_F - S_{M5}| \leq d \\ E_{OL}\{v_1, v_2, v_3, v_4\} + E_{OL6}, & |S_F - S_{M6}| \leq d \\ E_{OL}\{v_1, v_2, v_3, v_4\} + E_{OL7}, & |S_F - S_{M7}| \leq d \end{cases} \tag{8}$$

where $E_{OL}\{v_1, v_2, v_3, v_4\}$ presents the orientation distortion of the patch link. E_{OL4} , E_{OL5} , E_{OL6} and E_{OL7} are calculated by using the function terms in [15].

3.3.3 Scale Grid Line Distortion

There are three methods to compute the optimal scaling factor. Wang et al. [4] computed optical local scaling factor for each local region which scales different parts of an image differently and may change the proportions among diffract pats. Krähenbühl et al. [16] computed one global scaling factor. Niu et al. [15] computed the scaling factor by combining [16] and using different strategies to calculate the scaling factor for upsizing and downsizing. The idea of this method is computing optimal scaling factor by obtaining the minimal factor between w_f/w_s and h_f/h_s for upsizing and using the maximal factor between w_f/w_s and h_f/h_s for downsizing. We adopt the method [15].

After getting the scaling factor, we change the metrics from [2] to measure the scale distortion as follows:

$$\begin{cases} E_L\{v_1, v_2, v_3, v_4\} = (E_{LH} + E_{LV}) * S_m \\ E_{LH} = \left\| \dot{x}_4 - \dot{x}_1 - w_f * s_0 \right\|^2 + \left\| \dot{x}_3 - \dot{x}_2 - w_f * s_0 \right\|^2 \\ E_{LV} = \left\| \dot{y}_4 - \dot{y}_1 - h_f * s_0 \right\|^2 + \left\| \dot{y}_3 - \dot{y}_2 - h_f * s_0 \right\|^2 \end{cases} \quad (9)$$

where $E_L\{v_1, v_2, v_3, v_4\}$ denotes the scale distortion of the single mesh F, E_{LH} and E_{LV} present the horizontal and vertical scale distortion measure respectively, w_f and h_f are the width and height of the mesh F, s_0 is the scaling factor calculated by using [15].

Similar to the shape link and orientation link distortion, we apply constraints on scale links to keep the scale of the entire image information. The formulation of the measure metric is:

$$E_{LL}\{v_1, v_2, v_3, v_4\} = \begin{cases} E_{LL}\{v_1, v_2, v_3, v_4\} + E_{LL4}, & |S_F - S_{M4}| \leq d \\ E_{LL}\{v_1, v_2, v_3, v_4\} + E_{LL5}, & |S_F - S_{M5}| \leq d \\ E_{LL}\{v_1, v_2, v_3, v_4\} + E_{LL6}, & |S_F - S_{M6}| \leq d \\ E_{LL}\{v_1, v_2, v_3, v_4\} + E_{LL7}, & |S_F - S_{M7}| \leq d \end{cases} \quad (10)$$

$$\begin{cases} E_{LL4} = \left(\left\| \dot{x}_{44} - \dot{x}_1 - (w_F + w_{M4}) * s_0 \right\|^2 + \left\| \dot{x}_{43} - \dot{x}_2 - (w_F + w_{M4}) * s_0 \right\|^2 \right) * \left(\frac{S_F + S_{M4}}{2} \right) \\ E_{LL5} = \left(\left\| \dot{x}_{54} - \dot{x}_2 - (w_F + w_{M5}) * s_0 \right\|^2 + \left\| \dot{y}_{52} - \dot{y}_4 - (h_F + h_{M5}) \right\|^2 \right) * \left(\frac{S_F + S_{M5}}{2} \right) \\ E_{LL6} = \left(\left\| \dot{y}_{62} - \dot{y}_1 - (w_F + w_{M6}) * s_0 \right\|^2 + \left\| \dot{y}_{63} - \dot{y}_4 - (w_F + w_{M6}) * s_0 \right\|^2 \right) * \left(\frac{S_F + S_{M6}}{2} \right) \\ E_{LL7} = \left(\left\| \dot{x}_3 - \dot{x}_{71} - (w_F + w_{M7}) * s_0 \right\|^2 + \left\| \dot{y}_{72} - \dot{y}_1 - (h_F + h_{M7}) \right\|^2 \right) * \left(\frac{S_F + S_{M7}}{2} \right) \end{cases} \quad (11)$$

3.3.4 Boundary Constraints and Total Distortion

To preserve the completeness of the entire image, the boundary vertices of the target image are the ones of the original image. Suppose that the resolution of the original is

$w * h$ and the resolution of the output image is $w' * h'$, the boundary constraints are as follows:

$$\begin{cases} v'_{i,x} = 0, & \text{if } v_{i,x} = 0 \\ v'_{i,x} = w', & \text{if } v_{i,x} = w \\ v'_{i,y} = 0, & \text{if } v_{i,y} = 0 \\ v'_{i,y} = h', & \text{if } v_{i,y} = h \end{cases}$$

Base on the distortion measures defined above, we get the total distortion by combining all these distortion as following:

$$E = \sum_{F \in F} (p_s * E_s + p_o * E_o + p_L * E_L + p_{SL} * E_{SL} + p_{OL} * E_{OL} + p_{LL} * E_{LL}) \tag{12}$$

where $p_s, p_o, p_L, p_{SL}, p_{OL}$ and p_{LL} are the weight of single shape distortion, single orientation distortion, single scale distortion, shape link distortion, orientation link distortion and scale link distortion. We can transfer the different concern by change these weights.

4 Results and Discussion

We have implemented and tested our image retargeting method with matlab and C++ on a PC with 2.33 GHz Dual Core CPU and 2GB of memory. Our method is efficient because we formulate the minimization of total distortion as a linear system.

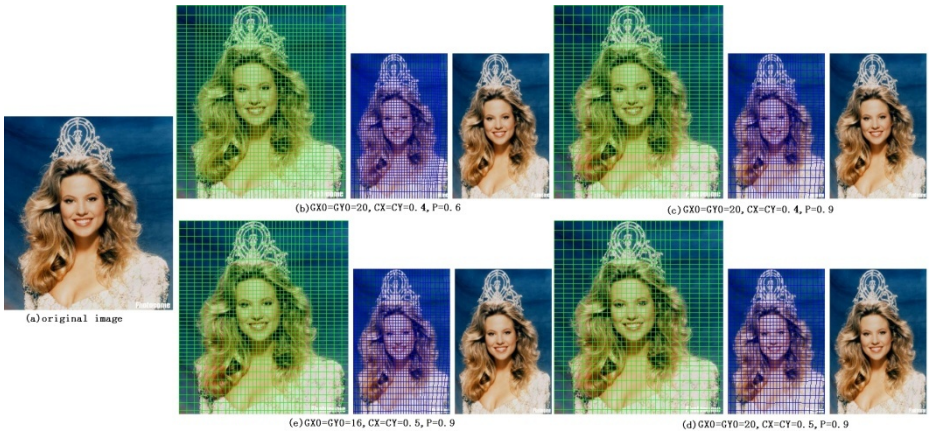


Fig. 4. An example that different factor value affects the result of constructing the mesh: (a) original image with 350*400; (b) constructing the mesh using $GX_0=GY_0=20, c_x=c_y=0.4, p=0.6$; (c) constructing the mesh using $GX_0=GY_0=20, c_x=c_y=0.4, p=0.9$; (d) constructing the mesh using $GX_0=GY_0=20, c_x=c_y=0.5, p=0.9$; (e) constructing the mesh using $GX_0=GY_0=16, c_x=c_y=0.5, p=0.9$; in(b, c, d, e), the first column is the mesh of the original image, the middle column is the mesh of the retargeted image, the right column is the targeted image



Fig. 5. Comparison of our results with those of scaling, seam carving (SC) [1] and optimal scale-and-stretch (OSS) [4]. The results of scaling, OSS [4] and our method tend to be smoother than those of seam carving. Notice the discontinuities in the ship, people, flowers, house and tower, which are caused by the pixels being removed. Compared with scaling and OSS [4], our method can preserve the aspect ratios of prominent features better.

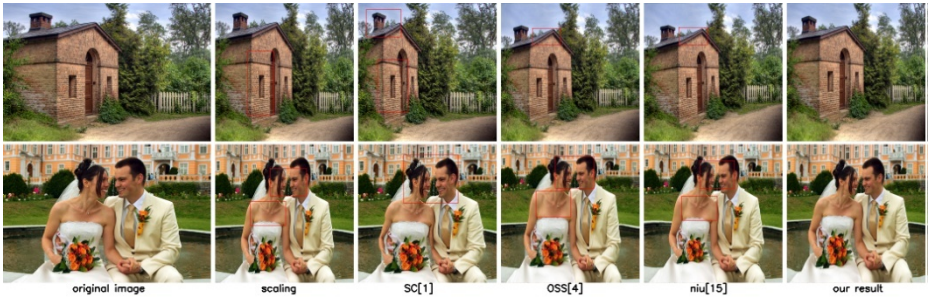


Fig. 6. Comparison of our results with those of scaling, SC[1], OSS[4] and Niu[15]. The results of scaling, OSS[4], Niu[15] and our method tend to be smoother than those of seam carving. Notice the discontinuities in the chest and the house roof and wall, which are due to the pixels being directly removed. Compared with scaling, OSS [4], Niu[15], our method can preserve the aspect ratios of prominent features better.

The computational cost relies on the parameter setting of the initialization and the factor, including p , GX_0 , GY_0 , c_x and c_y . The greater is p , the less is the range of the significant regions. The greater are GX_0 and GY_0 , the coarser are the meshes. The smaller are c_x and c_y , the finer are the meshes of the significant regions. We show these effects in Fig.4, but these effects are muted. In our experiments, we set $p=0.6$, $GX_0=20$, $GY_0=20$, $c_x=0.4$, $c_y=0.4$.

In Fig.5, we compare some results of homogeneous resizing, the seam carving (SC) [1] as implemented in Photoshop CS4, the optimized scale-and-stretch image resizing method (OSS) [4] and our method. It can be observed that the results of scaling, OSS [4] and our method produce smooth results, while seam carving produce noticeable discontinuity, especially in images containing structural objects. For example, the ship, people, tower, flowers, house since the pixels are directly removed. Comparing with the results of scaling and OSS [4], our method can preserve the aspect ratios of prominent objects. In Fig.6, we compare our results with Niu's [15]. We can see that our results preserve the important regions better, for example, the house roof is distortion in Niu [15], while it is straight in our result; the aspect ratio of the chest on the girl is changed, while preserved in our result.

5 Conclusion

In this paper, we present a novel image retargeting method based on non-uniform mesh warping, which adopts the different quadratic error metrics to minimize different distortion of the important regions and adapts the patch-linking scheme to apply constraints to the neighboring patches. Our method can preserve both the important regions and the global effect effectively. Moreover, since we link the neighboring meshes which are similar with each other, we save a large amount of time and increase efficiency. The experiment results show its effectiveness. Besides, our image retargeting method can be extended to process video by adding continuity constraint between adjacent frames and adding the motion information to the significance map.

Acknowledgements. We thank the anonymous reviewers for their insightful comments.

References

1. Avidan, S., Shamir, A.: Seam Carving for Content-Aware Image Resizing. *ACM Trans. Graph.* 26(3), 267–276 (2007)
2. Wolf, L., Guttman, M., Cohen-Or, D.: Non-homogeneous Content-driven Video-retargeting. In: *ICCV 2007: Proceedings of the Eleventh IEEE International Conference on Computer Vision*, pp. 1–6 (2007)
3. Rubinstein, M., Shamir, A., Avidan, S.: Improved Seam Carving for Video Retargeting. *ACM Trans. Graph.* 27(3), 1–9 (2008)
4. Wang, Y.-S., Tai, C.-L., Sorkine, O., Lee, T.-Y.: Optimized Scale-and-Stretch for Image Resizing. *ACM Trans. Graph.* 27(5), 1–8 (2008)
5. Wang, S.-F., Lai, S.-H.: Fast Structure-preserving Image Retargeting. In: *IEEE Intl. Conf. on Acoustics, Speech and Signal Processing*, pp. 1049–1052 (2009)
6. Zhang, G.-X., Cheng, M.-M., Hu, S.-M., Martin, R.R.: A Shape-Preserving Approach to Image Resizing. *Computer Graph Forum.* 28(7), 1897–1906 (2009)
7. Shreiner, D., Woo, M., Neider, J., Davis, T.: *The OpenGL programming guide—the official guide to learning OpenGL*, 5th edn. Addison-Wesley, Reading (2005)
8. Chen, L., Xie, X., Fan, X., Ma, W., Zhang, H., Zhou, H.: A visual attention model for adapting images on small displays. *Multimedia Syst.* 9(4), 353–364 (2003)
9. Itti, L., Koch, C., Niebur, E.: A model of saliency-based visual attention for rapid scene analysis. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 11(20), 1254–1259 (1998)
10. Dong, W., Zhou, N., Paul, J.-C., Zhang, X.: Optimized Image Resizing Using Seam Carving and Scaling. *ACM Trans. Graph.* 28(5), 1–10 (2009)
11. Gal, R., Sorkine, O., Cohen-Or, D.: Feature-aware texturing. In: *Proceedings of the 17th Eurographics Symposium on Rendering*, pp. 297–303 (2006)
12. Liu, L., Chen, R., Wolf, L., Cohen-Or, D.: Optimizing photo composition. *Compute. Graph. Forum* 29(2), 469–478 (2010)
13. Wang, D., Tian, X., Liang, Y., Qu, X.: Saliency-driven Shape Preservation for Image Resizing. *Journal of Information & Computational Science* 7(4), 807–812 (2010)
14. Viola, P., Jones, M.J.: Robust real-time face detection. *Int. J. Computer Vision* 57(2), 137–154 (2004)
15. Niu, Y., Liu, F., Li, X., Gleicher, M.: Image Resizing via non-homogeneous warping. *Multimed Tools Application* (2010)
16. Krähenbühl, P., Lang, M., Hornung, A., Gross, M.: A System for retargeting of streaming video. *ACM Trans. Graph.* 28(5), 1–10 (2009)
17. Harel, J., Koch, C., Perona, P.: Graph-based visual saliency. *Proc. of Neural Information Processing Systems* 19, 545–552 (2006)
18. Niu, Y., Liu, F., Li, X., Gleicher, M.: Warp propagation for video resizing. In: *CVPR 2010*, pp. 537–544 (2010)

Moving Edge Segment Matching for the Detection of Moving Object

Mahbub Murshed, Adin Ramirez, and Oksam Chae

Kyung Hee University, Yongin-si, Gyeonggi-do 446-701,
Republic of Korea

Abstract. We propose a segment based moving edge detection algorithm by building association from multi-frames of the scene. A statistical background model is used to segregate the moving segments that utilize shape and position information. Edge specific knowledge depending upon background environment is computed and thresholds are determined automatically. Statistical background model gives flexibility for matching background edges. Building association within the moving segments of multi-frame enhances the detection procedure by suppressing noisy detection of flickering segments that occurs frequently due to noise, illumination variation and reflectance in the scene. The representation of edge as edge segment allows us to incorporate this knowledge about the background environment. Experiments with noisy images under varying illumination changing situation demonstrates the robustness of the proposed method in comparison with existing edge pixel based moving object detection methods.

Keywords: Edge Segment, Moving Object Detection, Multi-frame based Edge matching, Statistical Background Model.

1 Introduction

The detection of moving object has been studied extensively due to the increasing demand in vision based applications like robotics, security, data compression, activity recognition system etc. Due to the simplicity of the detection procedure, background subtraction method for the detection of moving object has gained popularity. Here, current image is subtracted from the background image with a threshold. The automatic selection of this threshold value is very hard due to the nature of the application. Detecting moving object becomes more challenging where there are motion variations in the background. Moreover, a sudden noise spike or change in illumination or reflectance from other objects can have dramatic effect over the detection performance of a system. A comprehensive literature review on various moving object detection techniques can be found in [12] and [9]. There are two types of moving object detection approaches: the region based approach and the feature based approach. In the region based approach every pixel in the background is modeled. This is very sensitive since the intensity feature is very prone to illumination change. On the other hand, feature

based methods like edge, contour, curvature, corner, etc. tries to improve performance by utilizing feature strength, since features are less sensitive to illumination changing situation [12]. Among other feature based methods, edge based methods are popular since edge is more robust in the illumination change. Existing edge based moving object detection methods use edge differencing [11]. Kim and Hwang's method [10] uses edge pixel differencing algorithm using a static background. Thus the method gives scattered noise and cannot handle dynamic background. Dailey's method [5] computes background independent moving object by utilizing sequence image. But the method makes exact matching between edge pixels in consecutive frames. Thus the method brings out noise moreover it fails to detect slowly moving objects. Traditional edge based approaches also suffer from edge flickering. Edges flickers due to illumination variation, random noise, and reflectance from other objects. These flickering edge segments are not true moving edges. To cope with this difficulties, authors [13], tries to eliminate irrelevant edges by only selecting boundary edges. To solve edge inconsistency problems and to find object contour, authors used a multi-level canny edge map which is computationally very expensive. Even if they use multi-level canny edge map, to find a closed contour they needs to access image pixels directly, which is very noise sensitive. Edge segment based approach introduced by Hossain et al. [8], uses same chamfer [3] distance based matching method for both foreground and background edge segment. Since the characteristics of background edge segment are different from foreground edge segment, it is not suitable to use a common distance threshold for them. Moreover, their method cannot handle flickering edges that comes randomly due to the illumination reflectance of the other objects in the scene. We follow the approach proposed by Hossain et al. [8], but we have a separate evaluation technique for matching background edge and moving object edge. We use a statistical background model for matching every background edge segment separately as the motion variation of every background segment is not the same. Moreover, to overcome the problem of random flickering edges from the detected moving edges, we have used multi-frame based segment matching approach.

In the proposed method, edges from video frames are extracted using canny edge detector [4] and then we represent these edges as a structure of edge segments [1]. Here, a group of edge pixels form an edge segment and are processed together. An statistical background model adapts the motion variation of the background. Edge matching in multi-frame handles random flickering edges that evolve from the illumination reflectance of different objects by building associations within frames.

2 The Multi-frame Based Moving Edge Detection

The proposed method includes detection and verification of moving edge segments using a statistical background model for every input frame followed by building association within frames by matching detected moving edge segments.

For the detection of moving edge segments from a single frame, the system maintains two reference edge lists and a moving edge list. Static Background

Edge List (SBEL) is the first reference edge list that is generated by accumulating a number of training background edge image frames followed by thinning. Temporary Background Edge List (TBEL) is the other reference edge list. SBEL is a static list that is not updated but TBEL is updated at every frame. Moving Edge List (MEL) is made from the moving edges detected at current frame. Each edge segment in these lists has position, size and shape information. Moreover, TBEL edge segments have weight value with them.

Once moving edges are determined using a single input frame I_t , for every moving segment in that frame, we search for a matched correspondence with frame I_{t-1} . If the corresponding moving segment match is found within a distance threshold τ_d , the segment is placed in the output moving edge segment list for the frame I_t . The reason for this multi-frame match is straight forward; background model can eliminate background edges and edges from a stopped moving object from the scene but background model cannot handle flickering edges that comes occasionally by the illumination reflectance from background object or moving object. If an edge segment is true moving edge segment, it is more likely to come in consecutive frames where the flickering edges will not. The proposed moving edge detection method is given in Fig. 1.

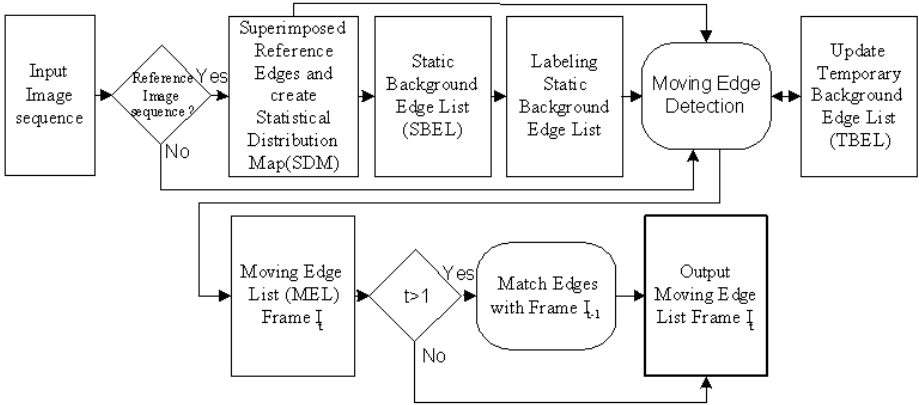


Fig. 1. The proposed moving edge detection method

2.1 The Statistical Background Model

Edges change their size and position within frames due to illumination change and noise. The amount of variation for different edge segment is different. Without considering this variation from the background, true moving edges cannot be detected.

Fig. 2 states the requirement to use statistical background model. Fig. 2(b) is made from the superimposition of twenty five reference edge lists. It is obvious that edges change their position and thus the edges in the superimposed edge image have thick lines. This thickness of the line is different for different background edge segment. Thus, in our proposed method, we treat every background



Fig. 2. (a) A sample reference background frame. (B) Edges from 25 superimposed background reference edge images.

segment individually. Using the statistical frequency accumulation information for each segment, we can restrict the search boundary that can also enhance the accuracy of matching as well as the speed.

The static background edge list (SBEL). Edges from training frames are extracted and we superimpose first N reference edge images using Eq. 1 and create accumulated reference edge image (AREI).

$$AREI^{(E,N)} = \sum_{p=1}^N \sum_{q=1}^z e_{p,q} \tag{1}$$

Here, $E = \{e\}$ is the edge map of an image, N is the total number of frames used, z is the number of edge segments on the p^{th} training image. After the accumulation, a smoothing operation is performed over the AREI. To make AREI independent of training sequence, we threshold AREI with $\tau\%$ of N . Here, we empirically found that $\tau = 30\%$ gives good result. Thus after thresholding, we produce Statistical Distribution Map (SDM) for the background. We create SBEL from the SDM by thinning SDM and extracting thin edge segments from the mid positions for every thick line. We then create edge segment labeling map for the extracted SBEL edge segment using SDM as shown in Fig. 3(d). The labeling map represents the search boundary for a candidate background edge segment during matching. We also have edge specific threshold for every SBEL

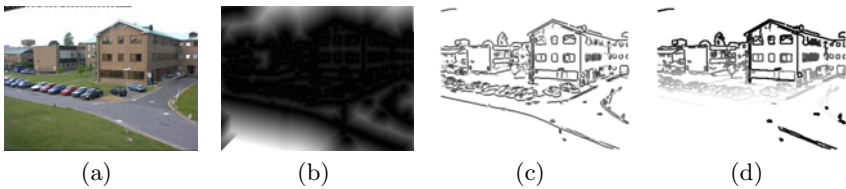


Fig. 3. Distance Map used in Hossain et al. method and the proposed method. (a) A sample reference frame. (b) CDM made from 50 training frames. (c) SDM made from 50 training frames. (d) Edge segment labeling map over the SDM.

segment by calculating the average accumulation score of each SBEL segment over the SDM.

Background edge segment matching. The background edge segment utilizes background edge segments statistic. Thus, background edge with high motion variation statistic will be matched with wider region and vice versa. For a given sample edge segment l , to determine whether it is a background edge, we compute average accumulation score SD by averaging the superimposed pixel positions over the SDM by using Eq. [2](#).

$$SD[l] = \left[\frac{1}{k} \sum_{i=1}^k SDM(l_i) \right] \quad (2)$$

Here, k is the number of edge point in the sample edge segment l , $SDM(l_i)$ is the edge point accumulation value in the background for sample edge point position l_i . From the labeled image of SBEL, we can find the candidate background segment directly. If no candidate background segment is found then the segment is a candidate moving edge segment. Otherwise, if the computed SD value is different from the corresponding background segment's average accumulation score by $T\%$, then the segment is also candidate moving edge segment. Otherwise, it is a background edge segment.

2.2 Multi-frame Based Moving Edge Matching

The problem at hand is to build partial segment matching between the candidate moving segments found at frame I_t and frame I_{t-1} . If moving segments from a moving object are detected correctly, there should be similarities between the detected moving edges in successive frames. As we have shown segments change their position within frames but the shape changes slowly. So if a moving segment has a significant portion of partial match in some consecutive frames, we can assume the segment as a true moving segment. A segment that does not show this shape consistency is surely a flickering edge that is generated due to illumination variation or reflectance from other object and hence we should discard these segments from the list of true moving segments. There are a number of curve matching solution that considers matching curve under affine transformation [2](#), the registration of 2D and 3D point set [7](#), distance based similarity measure based on multidimensional Hausdorff distance [18](#), all these methods give whole to whole curve matching solution with similarity measures. But our problem statement lies on the matching of whole to part matching problem with an index of the starting point of the match. [14](#) and [17](#) provides solution for whole to part matching problems but there method computes curvature points to reduce dimensionality. i.e. there method is suitable for those aligning problems where the problem statement needs to consider sharing and scaling as well. Also there methods are expensive to use in real time applications. Since we are matching moving edges in two consecutive frames, we can simplify our assumption that the matching curves can have some translation, small rotation, and

overall partial shape similarity. With this assumption, we need to match the edges considering partial shape match with translation and rotation only. The simple algorithm proposed in [15] best matches our interest. In their method two segments template segment and candidate segment are represented in slope angle-arclength space, or $\theta - a$ space by partitioning into segments of fixed arclength a_0 . In our case between a pair of segments to be matched, we assign the longer segment as candidate segment and shorter one as template segment. Matching is performed in $\theta - a$ space. During matching the template segment is moved along the a axis so that its centre is aligned with the centre of the candidate segment to which it is to be compared. The template segment is then shifted in the θ direction so that the mean θ value of the template segment has the same mean θ value as the image segment. This θ shift measures the average slope angle difference between them. The inverse of sum of the squares of these differences is used to measure the similarity between them. Finally, the location of the highest difference position along the a axis over the candidate segment is the position from where the two edge segments got match. For the details about the segment matching method please see [15].

2.3 Moving Edge Verification

Moving edges needs to be verified so that a stopped moving object is not detected as a moving object in future frames. A chamfer distance map is used to verify moving edge segments. A chamfer-3/4 distance map (CDM) [3] for the TBEL is created using Eq. 3.

$$CDM(i, j)^{(E)} = \min_{e \in E} |(i, j) - e| \tag{3}$$

Here, $E = \{e\}$ is the edge map of an image, i and j corresponds to row and column positions along the distance map. The distance value CD for any edge segment l can be computed using Eq. 4.

$$CD[l] = \frac{1}{3} \sqrt{\frac{1}{k} \sum_{i=1}^K CDM(l_i)^2} \tag{4}$$

Here, k is the number of edge point in the sample edge segment l , $CDM(l_i)$ is the i^{th} edge point distance value for the edge segment l .

To verify a moving edge segment, we create CDM for the high weighted segments from TBEL. Now the sample edge segment is placed over the CDM and distance value CD is calculated using equation Eq. 4. If CD is less than some threshold T_{CD} , then the segment is a non moving segment otherwise it is a moving segment.

2.4 Updating the TBEL

TBEL is constructed by adding the edges from MEL. If a moving edge is found in the same position in the next frame, the weight of that segment in TBEL is increased otherwise it is decreased. An edge segment will be dropped from TBEL if its weight reaches to zero.

3 Results and Analysis

Several experiments has been performed both in indoor and outdoor scene including parking lot, road scene, corridor. These images have background motion, illumination change, reflectance and noise. Our proposed method successfully detects almost all of the moving objects for the scene.

Fig. 4 shows the strength of our proposed method for varying illumination condition with noise. Fig. 4(a) shows a sample input frame No.890 of a street scene sequence. Four moving objects are (three people and a mini bus) present at the scene. Kim and Hwang [10] detects a lot of scattered edge pixels as shown in Fig. 4(b). The detection result for the Dailey and Cathey Method [5] is shown in Fig. 4(c). Hossain et al. method [8] can control camera movement in a limited scale but in their method the selection of a lower threshold results in matching mostly rigid background edges where as higher threshold increases false matching of moving edge as background edge. Moving object detection in our method utilizes movement statistic of every background edge segment effectively. Moreover, to eliminate flickering edges the proposed method tracks edge to edge matching record from multi-frame, thereby building association within moving edge

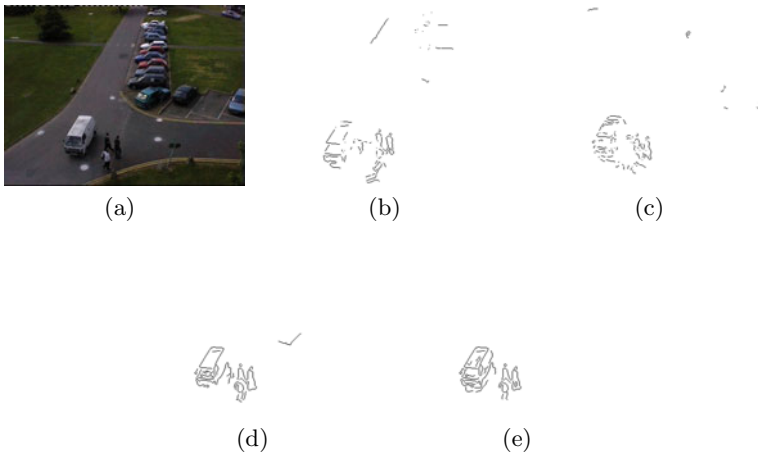


Fig. 4. (a) A sample input image frame No.890. (b) Detected moving edge image using Kim and Hwang’s method (c) Moving edge using Dailey and Cathey’s method (d) Detected moving edge segments proposed by Hossain at al. (e) Moving edge segments in the proposed method.

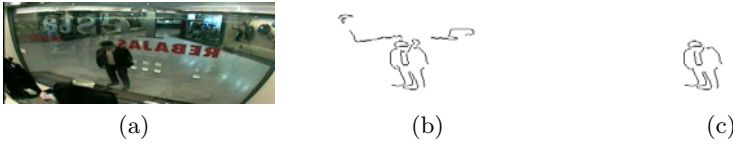


Fig. 5. (a) A sample image frame No.707. (b) Detected moving edge segments using the method proposed by Hossain et al. (c) Detected output moving edge segments using our proposed method.

segments. The detection output of our proposed method is given in fig. 4(e). Fig. 5 shows another example where we compared our method with Hossain et al. method since both of the methods have utilized edge segment structure. Using Hossain et al. method, due to illumination reflectance and noise, brings out flickering edge as moving edge that is found in Fig. 5(b). Our method, fig. 5(c), uses multi-frame edge segment matching, thus only true moving edge segments will have a good match. As a result, our detection output is more accurate and thus can significantly improve the performance of video surveillance based applications.

To evaluate the performance of the proposed system quantitatively, we compare the detected moving edge segments with the ground truth that is obtained manually. The metric used is based on two criteria: Precision and Recall and is defined in Eq. 5 and 6. Precision measures the accuracy of detecting moving edges while Recall computes the effectiveness of the extracted actual moving edge segments. The experimental result is shown in Table. 1.

$$Precision = \frac{Extracted\ moving\ edge\ pixels}{Total\ extracted\ edge\ pixels} \tag{5}$$

$$Recall = \frac{Extracted\ moving\ edge\ pixels}{Total\ actual\ moving\ edge\ pixels} \tag{6}$$

Table 1. Performance of the proposed moving edge detector

Dataset	Environment	Frames	Precision	Recall
1	outdoor	500	94%	88%
2	outdoor	400	98%	90%
3	indoor	500	93%	84%

For segmenting the moving objects, an efficient watershed based segmentation algorithm [16] can be used, where the region of interest (ROI) can be obtained by utilizing method [6].

4 Conclusion

This paper illustrates the suitability of using multi-frame based moving object detection method along with the statistical background model using segment based structure for the detection of moving object. Here, we utilized an efficient partial edge segment matching algorithm for inter-frame segment matching, a statistical background model for background edge segment matching and chamfer distance based matching for verifying moving edge segments from the scene. Our proposed method can eliminate flickering edges that comes occasionally. The example figures described in this paper clearly justifies the advantages of using statistical background model along with multi-frame based matching, which is highly efficient under illumination variation, reflection condition and background edge location changing situation. In our future work, we will incorporate edge contrast information with edge's side color distribution map for the matching and tracking of more sophisticated video surveillance based applications like intrusion detection, activity recognition etc.

References

1. Ahn, Y., Ahn, K., Chae, O.: Detection of moving objects edges to implement home security system in a wireless environment. In: Laganá, A., Gavrilova, M.L., Kumar, V., Mun, Y., Tan, C.J.K., Gervasi, O. (eds.) ICCSA 2004. LNCS, vol. 3043, pp. 1044–1051. Springer, Heidelberg (2004)
2. Bebis, G., Georgiopoulos, M., Lobo, N.D.V., Shah, M., Bebis, D.G.: Learning affine transformations. *Pattern Recognition* 32, 1783–1799 (1999)
3. Borgefors, G.: Hierarchical chamfer matching: a parametric edge matching algorithm. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 10(6), 849–865 (1988)
4. Canny, J.: A computational approach to edge detection. *IEEE Trans. Pattern Anal. Mach. Intell.* 8(6), 679–698 (1986)
5. Dailey, D.J., Cathey, F.W., Pumrin, S.: An algorithm to estimate mean traffic speed using uncalibrated cameras. *IEEE Transactions on Intelligent Transportation Systems* 1(2), 98–107 (2000)
6. Dewan, M.A.A., Hossain, M.J., Chae, O.: Background independent moving object segmentation for video surveillance. *IEICE Transactions* 92-B(2), 585–598 (2009)
7. Fitzgibbon, A.W.: Robust registration of 2d and 3d point sets. *Image and Vision Computing* 21(13-14), 1145–1153 (2003); *british Machine Vision Computing* 2001
8. Hossain, M.J., Dewan, M.A.A., Chae, O.: Moving Object Detection for Real Time Video Surveillance: An Edge Based Approach. *IEICE Trans. Commun.* E90-B(12), 3654–3664 (2007)
9. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. *IEEE Transactions on Systems, Man and Cybernetics* 34, 334–352 (2004)
10. Kim, C., Hwang, J.N.: Fast and automatic video object segmentation and tracking for content-based applications. *IEEE Trans. Circuits Syst. Video Techn.* 12(2), 122–129 (2002)

11. Makarov, A., Vesin, J.-M., Kunt, M.: Intrusion detection using extraction of moving edges. In: Proceedings of the 12th IAPR International Conference on Pattern Recognition, 1994. Conference A: Computer Vision and Image Processing, vol. 1, pp. 804–807 (October 1994)
12. Radke, R.J., Andra, S., Al-Kofahi, O., Roysam, B.: Image change detection algorithms: A systematic survey. *IEEE Transactions on Image Processing* 14, 294–307 (2005)
13. Roh, M.C., Kim, T.Y., Park, J., Lee, S.W.: Accurate object contour tracking based on boundary edge selection. *Pattern Recognition* 40(3), 931–943 (2007)
14. Sebastian, T.B., Klein, P.N., Kimia, B.B.: On aligning curves. *IEEE Trans. Pattern Anal. Mach. Intell.* 25(1), 116–125 (2003)
15. Turney, J.L., Mudge, T.N., Volz, R.A.: Recognizing partially occluded parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI* 7(4), 410–421 (1985)
16. Vincent, L., Soille, P.: Watersheds on digital spaces: An efficient algorithm based on immersion simulations. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 13(6), 583–598 (1991)
17. Wolfson, H.J.: On curve matching. *IEEE Trans. Pattern Anal. Mach. Intell.* 12(5), 483–489 (1990)
18. Yi, X., Camps, O.I.: Line-based recognition using a multidimensional hausdorff distance. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 21, 901–916 (1999)

Gauss-Laguerre Keypoints Extraction Using Fast Hermite Projection Method

Dmitry V. Sorokin, Maxim M. Mizotin, and Andrey S. Krylov

Laboratory of Mathematical Methods of Image Processing,
Faculty of Computational Mathematics and Cybernetics,
Lomonosov Moscow State University,
Moscow, Russia
{dsorokin,mizotin,kryl}@cs.msu.ru
<http://imaging.cs.msu.ru>

Abstract. Keypoints detection and descriptors construction method based on multiscale Gauss-Laguerre circular harmonic functions expansions is considered. Its efficient acceleration procedure is introduced. Two acceleration ideas are used. The first idea is based on the interconnection between Gauss-Laguerre circular harmonic functions system and 2D Hermite functions system. The further acceleration is based on the original fast Hermite projection method. The comparison tests with SIFT algorithm were performed. The proposed method can be additionally enhanced and optimized. Nevertheless even preliminary investigation showed promising results.

Keywords: keypoints extraction, Gauss-Laguerre circular harmonic functions, Hermite functions, fast Hermite projection method, image matching.

1 Introduction

The images keypoints extraction is one of the basic problems of low level image processing. Keypoints detection and parametrization is the initial step in tasks like stereo matching [1], object recognition [2], video indexing [3], panorama building and others. There are many approaches to the keypoints detection problem such as Harris corner detector [4], DoG approach presented by Lowe [5], the approach based on circular harmonic functions theory [6, 7], etc. The problem of keypoints descriptor construction is also widely presented in literature [4, 5, 8]. The invariance to a class of projective and photometric transformations is the target property of the descriptor construction algorithm. This property is crucial to obtain high matching rate across multiple views. As the majority of keypoints descriptors construction algorithms are computationally expensive, development of efficient computation algorithms becomes actual.

In this paper the keypoints detection and descriptors construction multiscale approach based on Gauss-Laguerre circular harmonic functions [6] is considered. The 2D Hermite projection-based fast algorithm for efficient exact keypoints

descriptors computation is proposed. The structure of the paper is the following: the first section is devoted to the Gauss-Laguerre keypoints extraction and its Hermite projection method acceleration, in the second section the fast Hermite projection method is considered and test results are described in the last section.

2 Gauss-Laguerre Keypoints

2.1 Gauss-Laguerre Keypoints Detection

Let us consider a family of complex orthonormal and polar separable functions:

$$\Psi(r, \gamma; \sigma) = \psi_n^{|\alpha|}(r^2/\sigma) e^{i\alpha\gamma} .$$

Their radial profiles are Laguerre functions:

$$\psi_n^\alpha(x) = \frac{1}{\sqrt{n! \Gamma(n + \alpha + 1)}} x^{\alpha/2} e^{-x/2} L_n^\alpha(x) ,$$

where $n = 0, 1, \dots; \alpha = 0, \pm 1, \pm 2, \dots$ and $L_n^\alpha(x)$ are Laguerre polynomials:

$$L_n^\alpha(x) = (-1)^n x^{-\alpha} e^x \frac{d}{dx^n} (x^{n+\alpha} e^{-x}) .$$

The Laguerre functions $\psi_n^\alpha(x)$ can be calculated using the following recurrence relations:

$$\begin{aligned} \psi_{n+1}^\alpha(x) &= \frac{(x - \alpha - 2n - 1)}{\sqrt{(n+1)(n+\alpha+1)}} \psi_n^\alpha(x) - \\ &\sqrt{\frac{n(n+\alpha)}{(n+1)(n+\alpha+1)}} \psi_{n-1}^\alpha(x) , \quad n = 0, 1, \dots, \\ \psi_0^\alpha(x) &= \frac{1}{\sqrt{\Gamma(\alpha+1)}} x^{\alpha/2} e^{-x/2} , \quad \psi_{-1}^\alpha(x) \equiv 0 . \end{aligned}$$

These functions $\Psi_n^\alpha(x)$, called Gauss-Laguerre circular harmonic functions (CHF), are referenced by integers n (referred by radial order) and α (referred by angular order). The real parts of $\Psi_n^\alpha(x)$ ($n = 0, 1, \dots, 4; \alpha = 1, 2, \dots, 5$) are illustrated in Fig. 1.

The Gauss-Laguerre CHF are self-steerable, i.e. they can be rotated by the angle θ using multiplication by the factor $e^{i\alpha\theta}$. They also keep their shape invariant under Fourier transformation. And they are suitable for multiscale and multicomponent image analysis [6], [9].

Let us consider an observed image $I(x, y)$ defined on the real plane R^2 . Due to the orthogonality of Ψ_n^α family the image $I(x, y)$ can be expanded in the neighborhood of the analysis point x_0, y_0 for fixed σ in Cartesian system as:

$$I(x + x_0, y + y_0) = \sum_{\alpha=-\infty}^{\infty} \sum_{n=0}^{\infty} g_{\alpha,n}(x_0, y_0; \sigma) \Psi_n^\alpha(\rho, \omega; \sigma) ,$$

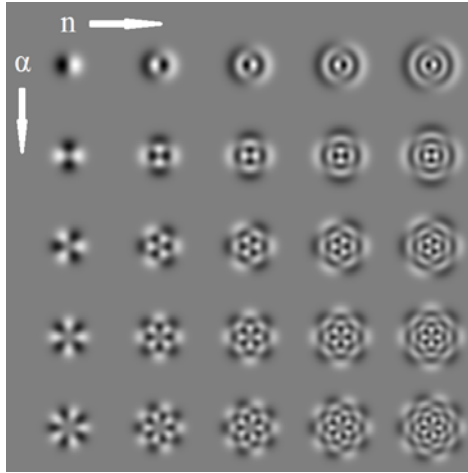


Fig. 1. The real part of Ψ_n^α ($n = 0, 1, \dots, 4; \alpha = 1, 2, \dots, 5$)

where

$$\rho = \sqrt{x^2 + y^2}, \quad \omega = \arctan\left(\frac{y}{x}\right),$$

and

$$g_{\alpha,n}(x_0, y_0; \sigma) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(x + x_0, y + y_0) \overline{\Psi_n^\alpha(\rho, \omega; \sigma)} dx dy .$$

Let us consider the keypoints detection algorithm introduced in [6]. Let σ be the scale parameter and $\sigma \in [2^{-s_{\max}}, 2^{s_{\max}}]$ discretized in $(2s_{\max} + 1)$ octaves where each octave contains N_s uniformly sampled scales. So the set of scales is defined as $\{\sigma_j\}$, where $j = 0, 1, \dots, 2N_s(2s_{\max} + 1) - 1$. Taking into account the Gauss-Laguerre CHF's property of being detectors for some image features (like edges, forks, crosses etc.), $n = 0, \alpha = 3, 4$ that corresponds to forks and crosses are considered. The set of $2N_s(2s_{\max} + 1)$ energy maps is defined as:

$$S(x, y; \sigma) = |g_{3,0}(x, y; \sigma)|^2 + |g_{4,0}(x, y; \sigma)|^2, \quad \sigma \in \{\sigma_j\} ,$$

referred as image scalogram. The scalogram is inspected by 3D sliding window $(5 \times 5 \times 3)$. The keypoints candidates $\overline{K} = (\overline{x}, \overline{y}; \overline{\sigma})$ are defined as the scalogram local maxima within the window. Here $(\overline{x}, \overline{y})$ is the keypoint coordinate and $\overline{\sigma}$ is the keypoint reference scale. So the image keypoints set is $\{\overline{K}\}$. This set is reduced by rejecting those keypoints \overline{K} which have the same position $(\overline{x}, \overline{y})$ for more than two reference scales. And, finally, the keypoints \overline{K} with energy value $S(\overline{x}, \overline{y}; \overline{\sigma})$ less than a selected threshold are omitted:

$$S(\overline{x}, \overline{y}; \overline{\sigma}) < T \cdot \max_{x,y} (S(x, y; \overline{\sigma})) . \tag{1}$$

$T \in [0, 1]$ is adjustable parameter and it is used to control the number of detected keypoints. In our tests T was set to get about 1000 keypoints in each image of the pair.

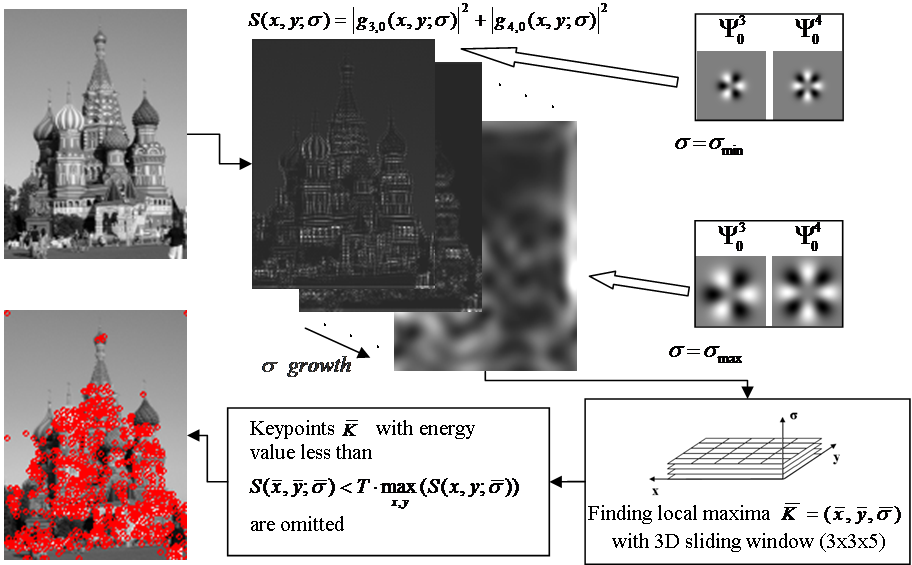


Fig. 2. The flowchart of the keypoints detection process

The flowchart of the keypoints detection algorithm is illustrated in Fig. 2

2.2 Gauss-Laguerre Keypoints Descriptors

The Gauss-Laguerre keypoints descriptors construction algorithm was first proposed in [6]. Each keypoint $\bar{K} = (\bar{x}, \bar{y}; \bar{\sigma})$ is associated to a local descriptor $\bar{\chi} = \{\bar{\chi}(n, \alpha, j)\}$. This is a complex-valued vector consisted of local image projections to a set of Gauss-Laguerre CHF's Ψ_n^α at $2j_{\max}$ scales neighbor to the keypoint \bar{K} reference scale $\bar{\sigma}$. The $\bar{\chi}$ elements are defined as:

$$\bar{\chi}(n, \alpha, j) = \frac{g_{\alpha,n}(x, y; \sigma_j) \cdot e^{-i\alpha\theta_j}}{\|g_{\alpha,n}(x, y; \sigma_j) \cdot e^{-i\alpha\theta_j}\|},$$

$$n = 0, \dots, n_{\max}, \alpha = 1, \dots, \alpha_{\max}, j = -j_{\max}, \dots, j_{\max},$$

where σ_j is the j -th scale following $\bar{\sigma}$ if $j > 0$, or preceding $\bar{\sigma}$ if $j < 0$ in the discretized scale space. The normalization makes descriptor invariant to the contrast changes. The phase shift $e^{-i\alpha\theta_j}$ is used to make the descriptors invariant to the keypoint pattern orientation, where

$$\theta_j = \arg(g_{1,0}(\bar{x}, \bar{y}; \sigma_j)).$$

The matching performance of this technique was demonstrated in [6] in comparison with SIFT algorithm. It was found in [6] that Gauss-Laguerre keypoints extraction method matching results overcome SIFT algorithm results in the case of rotation, scale and translation transformation of images. Nevertheless the computational cost of the algorithm is high.

2.3 Descriptors Computation Using 2D Hermite Functions Expansion

The 2D Hermite functions $\Phi_{m,n}(x, y; \sigma)$ form the complete orthonormal system in L_2 space and can be defined as:

$$\Phi_{m,n}(x, y; \sigma) = \frac{1}{\sigma} \phi_m\left(\frac{x}{\sigma}\right) \phi_n\left(\frac{y}{\sigma}\right), \phi_n(x) = \frac{1}{\sqrt{2^n n! \sqrt{\pi}}} e^{-\frac{x^2}{2}} H_n(x), \quad (2)$$

where $n = 0, 1, 2, \dots$ and $H_n(x)$ are Hermite polynomials:

$$H_n(x) = (-1)^n e^{x^2} \frac{d}{dx^n} (e^{-x^2}).$$

The Hermite functions $\phi_n(x)$ can be calculated using the following recurrence relations:

$$\begin{aligned} \phi_n(x) &= x \sqrt{\frac{2}{n}} \phi_{n-1}(x) - \sqrt{\frac{n-1}{n}} \phi_{n-2}(x), \quad n = 2, 3, \dots, \\ \phi_0(x) &= \frac{1}{\sqrt[4]{\pi}} e^{-\frac{x^2}{2}}, \quad \phi_1(x) = \frac{\sqrt{2}x}{\sqrt[4]{\pi}} e^{-\frac{x^2}{2}}. \end{aligned}$$

The 2D Hermite image $I(x, y)$ expansion in the analysis point x_0, y_0 for fixed σ can be defined as:

$$I(x + x_0, y + y_0) = \sum_{m=0}^{\infty} \sum_{n=0}^{\infty} h_{m,n}(x_0, y_0; \sigma) \Phi_{m,n}(x, y; \sigma),$$

where

$$h_{m,n}(x_0, y_0; \sigma) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} I(x + x_0, y + y_0) \Phi_{m,n}(x, y; \sigma) dx dy. \quad (3)$$

As one can see from (2), $\Phi_{m,n}(x, y; \sigma)$ functions are Cartesian separable, so the computation of (3) can be performed as:

$$\bar{h}_{m,n}(x_0, y + y_0; \sigma) = \int_{-\infty}^{\infty} I(x + x_0, y + y_0) \phi_m\left(\frac{x}{\sigma}\right) dx, \quad (4)$$

for every fixed y and after that

$$h_{m,n}(x_0, y_0; \sigma) = \frac{1}{\sigma} \int_{-\infty}^{\infty} \bar{h}_{m,n}(x_0, y_0 + y; \sigma) \phi_n\left(\frac{y}{\sigma}\right) dy. \quad (5)$$

The idea of using the interconnection of 2D Hermite functions and Gauss-Laguerre CHF was first introduced in [10]. Any Gauss-Laguerre CHF can be represented as the linear combination of 2D Hermite functions [11], [12] (the example of connection between Gauss-Laguerre CHF and 2D Hermite functions

is illustrated in Fig. 3). So the corresponding coefficients $g_{\alpha,n}$ and $h_{m,n}$ of image expansion to the sets of these functions are connected with the same relation. The formulae and more detailed description of the interconnection can be found in [10].

Using the separability of $\Phi_{m,n}$ functions and interconnection between $\Phi_{m,n}$ and Ψ_n^α functions the number of operations for $g_{\alpha,n}$ computation can be reduced up to several times.

To suppress the descriptor changes due to the brightness changes we introduce the following step. Before expanding the image in keypoint neighborhood into the set of Gauss-Laguerre CHF's the average value of keypoints boundary pixels intensity is subtracted from keypoint neighborhood image intensity values.

Further acceleration can be achieved using fast Hermite projection method to compute coefficients $\bar{h}_{m,n}$ and $h_{m,n}$ in 1D expansions (4), (5).

$\Phi_{3,0}$	$\Phi_{2,1}$	$\Phi_{1,2}$	$\Phi_{0,3}$	Re	Im	
						$\Psi_{0,-3}$
$\frac{1}{2\sqrt{2}}$	$-\frac{3}{2\sqrt{2}}i$	$-\frac{3}{2\sqrt{2}}$	$\frac{1}{2\sqrt{2}}i$			$\Psi_{1,-1}$
$\frac{3}{2\sqrt{2}}$	$-\frac{1}{2\sqrt{2}}i$	$\frac{1}{2\sqrt{2}}$	$-\frac{3}{2\sqrt{2}}i$			$\Psi_{1,1}$
$\frac{3}{2\sqrt{2}}$	$\frac{1}{2\sqrt{2}}i$	$\frac{1}{2\sqrt{2}}$	$\frac{3}{2\sqrt{2}}i$			$\Psi_{0,3}$
$\frac{1}{2\sqrt{2}}$	$\frac{3}{2\sqrt{2}}i$	$-\frac{3}{2\sqrt{2}}$	$\frac{1}{2\sqrt{2}}i$			

Fig. 3. An example of relation between Gauss-Laguerre CHF's and 2D Hermite functions. The matrix M_4 of connection between subset of 4 Gauss-Laguerre CHF's and 2D Hermite functions is illustrated. $[\Psi_{0,-3} \ \Psi_{1,-1} \ \Psi_{1,1} \ \Psi_{0,3}]^T = M_4 \cdot [\Phi_{3,0} \ \Phi_{2,1} \ \Phi_{1,2} \ \Phi_{0,3}]^T$.

3 Fast Hermite Projection Method

In common case 1D Hermite projection method is defined as:

$$f(x) = \sum_{m=0}^{\infty} c_m \phi_m(x)$$

where $\phi_m(x)$ are 1D Hermite functions, c_m are Hermite coefficients:

$$c_m = \int_{-\infty}^{\infty} f(x) \phi_m(x) dx . \tag{6}$$

Each coefficient in (6) can be rewritten through Hermite polynomials as follows:

$$c_m = \frac{1}{\beta_m} \int_{-\infty}^{\infty} e^{-x^2} \left(f(x)e^{\frac{x^2}{2}} \right) H_m(x) dx ,$$

where $H_m(x)$ is Hermite polynomial, β_m is Hermite normalization constant:

$$\beta_m = \sqrt{2^m m! \sqrt{\pi}} .$$

This integral can be approximated by Gauss-Hermite quadrature [13]:

$$c_m = \frac{1}{\beta_m} \int_{-\infty}^{\infty} e^{-x^2} \left(f(x)e^{\frac{x^2}{2}} \right) H_m(x) dx \approx \frac{1}{\beta_m} \sum_{k=1}^N A_k \left(f(x_k)e^{\frac{x_k^2}{2}} \right) H_m(x_k) ,$$

where x_k – Hermite polynomials $H_N(x)$ zeros, A_k – associated weights:

$$A_k = \frac{2^{N-1} N! \sqrt{\pi}}{N^2 H_{N-1}^2(x_k)} . \tag{7}$$

Computation cost and precision loss of these associated weights increase with the increase of N [14]. This problem can be solved by replacement of Hermite polynomials by Hermite functions in (7) [14]. After simplification the following formula can be obtained:

$$c_m \approx \frac{1}{N} \sum_{k=1}^N \mu_{N-1}^m(x_k) f(x_k) ,$$

where $\mu_{N-1}^m(x_k)$ is an array of associated constants:

$$\mu_{N-1}^m(x_k) = \frac{\phi_m(x_k)}{\phi_{N-1}^n(x_k)} .$$

More details on fast Hermite projection method can be found in [14].

Keypoints descriptors elements computation can be even more accelerated using fast Hermite projection method to calculate $\bar{h}_{m,n}$ and $h_{m,n}$ in (4) and (5). However fast Hermite Projection method is lossy. So this acceleration brings in some error to the Gauss-Laguerre image expansion coefficients $g_{\alpha,n}$ and as a consequence keypoints descriptors elements.

4 Results

Proposed keypoints extraction algorithm has been tested on the images selected from the dataset freely available on the web, which provides the image and the relating homographies sequences (<http://www.robots.ox.ac.uk/~vgg/research/affine/>).

Typical values of achieved acceleration of initial descriptor construction algorithm are demonstrated in Table 1. The threshold T in keypoints detection (1) was set for each image independently to get about 1000 keypoints per image. The values of descriptors construction parameters were $n = 5$, $\alpha = 5$, $j_{\max} = 2$. Fast Hermite projection method was applied for keypoints with reference scale $\sigma > 5$. This value was chosen experimentally to get optimal balance between acceleration and approximation errors.

Table 1. Method acceleration results

Image name	2D Hermite separability acceleration	Fast Hermite projection method acceleration	Overall acceleration
boat1	3.77	1.42	5.36
boat2	3.80	1.45	5.50
boat3	3.82	1.39	5.31
graf1	1.44	3.22	4.67
graf2	1.49	3.25	4.85
graf3	1.49	3.37	5.02

The complete comparison of computational cost of proposed acceleration of Gauss-Laguerre descriptors construction algorithm and SIFT descriptors construction algorithm (5) is not given in this paper due to the fact that current implementation of Gauss-Laguerre keypoints descriptors construction algorithm is not optimized. Current implementation of the Gauss-Laguerre algorithm with the fast Hermite projection method acceleration is ~ 10.5 times slower than implementation of SIFT which is freely available on the web (<http://www.robots.ox.ac.uk/~vgg/research/affine/>).

The proposed method was compared in precision-recall (8) with SIFT keypoints descriptors construction algorithm. Descriptors were constructed for the same set of keypoints (5) selected with Gauss-Laguerre keypoints detection algorithm. Threshold T was identical for both pair images and its value was set to get at least 1000 keypoints in both images. The values of Gauss-Laguerre descriptors construction parameters were $n = 5$, $\alpha = 5$, $j_{\max} = 2$. Fast Hermite projection method was applied for keypoints with reference scale $\sigma > 5$. Different recall values were obtained changing the nearest neighbor distance ratio parameter in descriptors matching procedure proposed by Lowe (5).

Typical results are illustrated in Fig. 4, 5. The proposed method needs additional enhancement and optimization. Nevertheless even preliminary investigation showed promising results.

In Fig. 4 the results for graf1-graf2 image pair are given. This pair corresponds to points of view changing transformation. The obtained results show that Gauss-Laguerre descriptors and fast modification of Gauss-Laguerre descriptors perform better matching than SIFT descriptors for the same level of recall. However SIFT descriptors allow to reach the higher level of recall.

In Fig. 5 the results for boat1-boat2 image pair are given. This pair corresponds to rotation and zoom transformations. The obtained results show that Gauss-Laguerre descriptors performs better matching than SIFT descriptors for

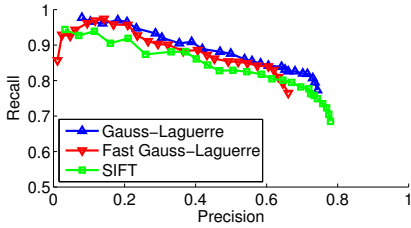


Fig. 4. Precision-Recall graph with different descriptors for graf1-graf2 image pair

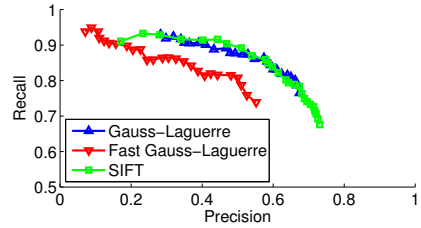


Fig. 5. Precision-Recall graph with different descriptors for boat1-boat2 image pair

some levels of recall, but SIFT descriptors outperforms proposed descriptors in the area of high values of recall. Hermite projection based Gauss-Laguerre descriptors demonstrate less level of both recall and precision than SIFT and Gauss-Laguerre descriptors.

5 Conclusion

The efficient computation technique of Gauss-Laguerre keypoints descriptors using both the interconnection between Gauss-Laguerre circular harmonic functions and 2D Hermite functions and fast Hermite projection method have been proposed. The preliminary test results look promising. Nevertheless the tests showed that proposed descriptors are not fully invariant to brightness and contrast changes. Future work will include investigation in the field of brightness and contrast invariance of the descriptors and further improvement of Gauss-Laguerre keypoints detection algorithm.

Acknowledgments. The work was supported by RFBR grant 10-01-00535-a.

References

1. Schaffalitzky, F., Zisserman, A.: Multi-view Matching for Unordered Image Sets. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) ECCV 2002. LNCS, vol. 2350, pp. 414–431. Springer, Heidelberg (2002)
2. Tuytelaars, T., Ferrari, V., Van Gool, L.: Simultaneous object recognition and segmentation from single or multiple model views. *Int. J. of Computer Vision* 67(2), 159–188 (2006)
3. Morand, C., Benois-Pineau, J., Domenger, J.-P., Zepeda, J., Kijak, E., Guillemot, C.: Scalable object-based video retrieval in HD video databases. *J. Signal Processing: Image Communication* 25(6), 450–465 (2010)
4. Harris, C.G., Stephens, M.: A combined corner and edge detector. In: 4th Alvey Vision Conf. Manchester, pp. 147–151 (1988)
5. Lowe, D.: Distinctive image features from scale-invariant keypoints. *Int. J. of Computer Vision* 60(2), 91–110 (2004)

6. Sorgi, L., Cimminiello, N., Neri, A.: Keypoints Selection in the Gauss Laguerre Transformed Domain. In: BMVC 2006, pp. 133–142 (2006)
7. Hse, H., Newton, A.R.: Sketched symbol recognition using zernike moments. In: ICPR 2004, pp. 367–370 (2004)
8. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. *IEEE Trans. on PAMI* 27(10), 1615–1630 (2005)
9. Jacovitti, G., Neri, A.: Multiresolution circular harmonic decomposition. *IEEE Trans. Signal Processing* 48(11), 3243–3247 (2000)
10. Sorokin, D.V., Krylov, A.S.: Fast Gauss-Laguerre Keypoints Extraction using 2D Hermite Functions. In: PRIA-10-2010, pp. 339–342 (2010)
11. Zauderer, E.: Complex argument Hermite-Gaussian and Laguerre-Gaussian beams. *J. Opt. Soc. Amer. A* 3(4), 465–469 (1986)
12. Di Claudio, E.D., Jacovitti, G., Laurenti, A.: Maximum Likelihood Orientation Estimation of 1-D Patterns in Laguerre-Gauss Subspaces. *IEEE Tran. Image Processing* 19(5), 1113–1125 (2010)
13. Krylov, V.I.: *Approximate Calculation of Integrals*. Macmillan Press, New York (1962)
14. Krylov, A., Korchagin, D.: Fast Hermite Projection Method. In: Campilho, A., Kamel, M.S. (eds.) ICIAR 2006. LNCS, vol. 4141, pp. 329–338. Springer, Heidelberg (2006)

Re-identification of Visual Targets in Camera Networks: A Comparison of Techniques

Dario Figueira and Alexandre Bernardino

Institute for Systems and Robotics,
Instituto Superior Técnico,
1049-001 Lisboa, Portugal
{dfigueira,alex}@isr.ist.utl.pt

Abstract. In this paper we address the problem of re-identification of people: given a camera network with non-overlapping fields of view, we study the problem of how to correctly pair detections in different cameras (one to many problem, search for similar cases) or match detections to a database of individuals (one to one, search for best match case). We propose a novel color histogram based features which increases the re-identification rate. Furthermore we evaluate five different classifiers: three fixed distance metrics, one learned distance metric and a classifier based on sparse representation, novel to the field of re-identification. A new database alongside with the matlab code produced are made available on request.

Keywords: Re-Identification, distance metrics, pattern recognition, visual surveillance, camera network.

1 Introduction

Re-identification is still an open problem in computer vision. The enormous possible variations from camera to camera in illumination, pose, color or all of those combined, introduce large appearance changes on the people detected, which make the problem very difficult to overcome.

Re-identification denotes the problem of given multiple cameras, and several people passing in front of several cameras, to determine which person detected in camera X corresponds to the person detected in camera Y.

There are a few works in the literature addressing the problem of re-identification in camera networks. [9] uses the bag-of-visual-words approach, clustering SIFT [8] features into “words”, and using those “words” to describe the detections, in a one to one approach to re-identification in a shopping center environment. This approach is of interest because it merges the “very high detail/specificity” of a SIFT feature with the generalization power of a cluster (a “word”). [5] uses SURF [1] features also in a one to one approach to re-identification in a shopping center environment (CAVIAR database [4]). SURF’s are extracted from the image’s hessian space, as opposed to SIFT’s features that are extracted from the image’s laplacian space. SURF’s are also much faster to

¹ <http://groups.inf.ed.ac.uk/vision/CAVIAR/CAVIARDATA1/>

be computed, which is of note since SIFT's major drawback is its heavy computation time. Both works use a voting classifier that is already standard when using such features. Also of note is [10] for its simple features, histograms, used also in a one to one approach, to re-identify people in similar poses walking inside a train, testing different histograms and normalizations to cope with greatly varying illumination. They employ dimensionality reduction and nearest-neighbor for a classifier. [6] also uses simple histogram features but takes advantage of the appearance and temporal relationship between cameras.

Our work is similar to [6, 10]'s in term of used features (histograms), although we enrich them by considering the upper and lower body parts separately. A method to detect the waist of a person is proposed, which allows the separate representation of the colors on the upper and lower body parts. This aspect highly distinguishes our work and significantly improves the re-identification results. Also we test additional metric distances including one learned from the data. Furthermore we apply a sparse base classifier, successful in the domain of face recognition, to the re-identification problem.

In this work we not only consider the one-to-one approach, where given a person detection we want to recognize it in a database, but also a one-to-many approach, where someone is trying to find similar matches in the system to a given person's image.

We produced an indoors dataset where we evaluate several techniques and the effects of our enriched feature, in both approaches.

In the next Section we define the problem. In Section 3 we propose our new color based feature. In Section 4 we list the metrics reviewed. In Section 5 we describe the data used and show the experimental results. Finally we conclude in the last Section.

2 Problem Definition

In this Section we define our problem, while the approaches are described in the following sub-sections.

Figure 1 depicts the environment: A network of fixed cameras with non-overlapping fields of view, where people appear more than once, are detected, tracked while in camera view, and then re-identified between different cameras. While a person is in view, we track it, and extract the following feature from it:

1. Detect the person and extract its pixels with [2]'s background subtraction;
2. Normalize the color of the detection pixels with greyworld normalization [10];
3. Divide the image in two by the waist (detailed in Section 3);
4. Compute the color histograms of each part, and unit normalize them;
5. Compute the mean (μ) and covariance (Σ) for all histograms in the track sequence to obtain one point per track sequence.

Therefor each track becomes an averaged histogram feature, which will then be a point (x or y) in the following formulations.

2.1 One to One Problem - Recognize - "Who is this person?"

This is the standard re-identification approach, used in [9, 5, 10], and applied in real world situations, where a surveillance operator picks out a person detection and asks



Fig. 1. Two camera views, with several detections, waist-separated, and two tracks. A correct re-identification would cluster them together or label them as the same individual from a database.

the system to recognize it. This approach is of particular interest in scenarios where you have a controlled entrance to the system. In such an entrance we can easily insert the incoming individuals into a database along with their identifications.

So in this case we always have a training set, a dataset of labeled tracks to start with. Given a test sample we compute the best match to the classes in the training set for such test sample.

2.2 One to Many Problem - Search - “Where Was This Person?”

In this work we also consider the one to many approach, where a surveillance operator sees a person in a camera (identified or not), and asks the system to show him all related detections, in all cameras.

Given the track points, we compute the distances from all to all, then solve the binary classification problem of determining, given an appropriate distance metric, if a pair of tracks belongs to a single person, or come from different people. By varying the

distance threshold that determines if such a distance is small enough to belong to a pair of tracks from a single person, we output a ROC curve. Thus examining the ability for each re-identification technique to correctly cluster points.

We also study the advantage of learning a metric from data versus a standard distance metric.

3 Person Representation

Simple color histograms have been used as the appearance features in tracking across cameras [6,10]. We enrich such features by dividing the person histogram in two parts, one above and one below the waist. We define the waist as the point that maximizes the Euclidean distance between the upper part histogram and lower part histogram. After computing the integral histogram, we do a vertical search for the waist, limiting the search to an area around the middle of the image.

- Compute Vertical Integral Histogram
 - Compute the histogram of the first horizontal line of the image; compute the histogram of the first two lines of the image; ...; compute the histogram of the whole image.
- Search for Point that Maximizes Distance Between Upper and Lower Body Parts
 - Compare line by line, the upper and lower histograms; Plot the varying distance; Find maximum.
 - Limit search to window between 35 and 60% of the image, counting from the top (maximum and minimum empirical values of position of the waist found during the manual labeling of the dataset).

4 Re-identification Techniques

In this section we describe the methods used in the automatic re-identification system. We compared three distance metrics: Euclidean; Bhattacharya; and diffusion distance [7] with one linear metric learning method, and one recent classification method [11] developed in the face recognition field. Each track is represented by \mathbf{x} or \mathbf{y} , as stated in Section 2.

We consider the following metrics to compute distance between the histograms.

Euclidean. A simple nearest neighbor distance. $d_E(\mathbf{x}, \mathbf{y}) = \|\mathbf{x} - \mathbf{y}\|$.

Bhattacharya. Modified Bhattacharya coefficient [3], a common choice for measuring distance between histograms.

$$d_{BHATT}(\mathbf{x}, \mathbf{y}) = \sqrt{1 - \sum_{i=1}^m \sqrt{x_i y_i}}$$

Diffusion Distance. Given we use histograms for features, we looked for alternative ways to measure distances between histograms. The Earth's Mover Distance (EMD) is popular, and the Diffusion Distance [7] has been shown to have equal or better results than EMD, with the added benefit of faster computation.

$$\begin{aligned}
 d_{DIFF}(\mathbf{x}, \mathbf{y}) &= \sum_{l=0}^L |d_l| \\
 d_0 &= \mathbf{x} - \mathbf{y} \\
 d_l &= [d_{l-1} * \phi(d_{l-1})] \downarrow_2 \quad l = 1, \dots, L \\
 \downarrow_2 &: \text{downsample to half-size} \\
 \phi(\cdot) &: \text{gaussian filter}
 \end{aligned}$$

Metric Learning. The work of Xing in [12] linearly learns a metric of the following form:

$$d_{ML}(x, y) = \|x - y\|_A = \sqrt{(x - y)^T A (x - y)},$$

by solving the following optimization problem

$$\begin{aligned}
 \arg \max_A \quad & \sum_{(i,j) \in D} \|x_i - x_j\|_A = \sum_{(i,j) \in D} \sqrt{d^{ijT} A d^{ij}} \\
 \text{s.t.} \quad & \sum_{(i,j) \in S} \|x_i - x_j\|_A^2 = \sum_{(i,j) \in S} d^{ijT} A d^{ij} \leq t \\
 & A \geq 0
 \end{aligned}$$

where t is a scalar, S and D are square binary matrixes that represent the similar (from a same person) and dissimilar (from different people) training pair sets (in S , one if a pair is similar, and zero otherwise. Likewise in D). We implemented this optimization problem in Matlab with the CVX’s optimization toolbox². A training set of similar and dissimilar pairs is required.

Sparse Recognition Classifier. Sparsity has been widely used in signal processing for reconstruction [4]. Here we apply it to recognition, in the form of a re-identification problem. Simply put, given a test sample \mathbf{y} , we solve the optimization problem

$$\begin{aligned}
 \arg \min_{[\mathbf{i} \ \mathbf{e}]} \quad & \|[\mathbf{i} \ \mathbf{e}]\|_1 \\
 \text{s.t.} \quad & [A \ I] [\mathbf{i} \ \mathbf{e}]^T = \mathbf{y} \\
 & A = [x_1 \ \dots \ x_T]
 \end{aligned}$$

where in the columns of A are the T training samples (*i.e.*, three random histogram vectors from three random detections from each person to be recognized). The reasoning behind this formulation is that we wish to choose from A which training class \mathbf{y} belongs to. This information will be encoded in the indicator vector \mathbf{i} , while errors will be explicitly modeled in \mathbf{e} . Moreover, by minimizing the l_1 norm of $[\mathbf{i}, \mathbf{e}]^T$, and if the true solution is sparse, the l_1 norm minimization will output the same result as the l_0 norm [4], the sparsest solution. \mathbf{i} will then be mostly zero with few large entries in the correct training set entries.

This idea has first been put forth in the field of face recognition by [11].

² <http://cvxr.com/cvx/>

5 Results

In this Section, first we describe the experimental setup where the datasets were taken. We present our results in sub-sections 5.3 and 5.4, we validate our proposed feature, and we discuss the results.

5.1 Experimental Setup

We produced a database of images, extracted from an indoors camera network, completely hand labeled for ground truth. In Figure 2 we show some samples of the varying camera views from the dataset.

Indoor Dataset:

- 17388 detections;
- 275 tracks;
- 26 people (5 of which only have one track).
- 10 fixed cameras, with non-overlapping fields of view.
- Average 67 detections per track, Maximum 205 detections in a track.

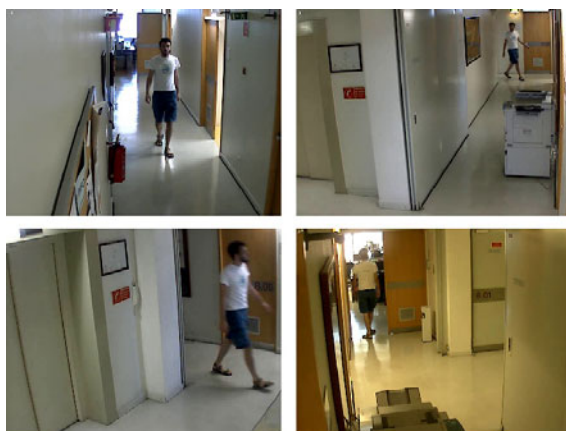


Fig. 2. Indoor Dataset: Preview of some camera views

For all experiments we computed the detections and features as described in Section 2. These combine graycolor normalization and division by the waist. Figure 3 supports the choices made.

5.2 Training

For the metric learning training and testing we used 5-fold cross validation. For Sparse Recognition Classifier training, we picked 3 detection points per person to form the training class matrix (A).

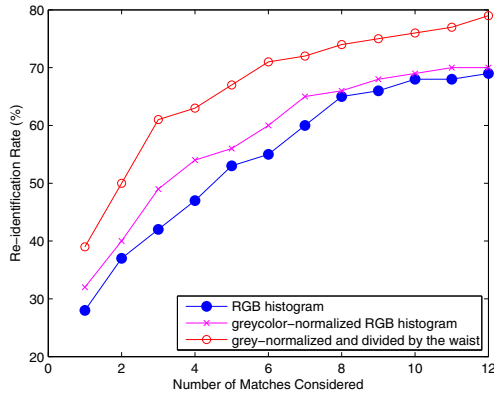


Fig. 3. Cumulative Matching Characteristic curve of several feature combinations with Euclidean metric. Re-identification results for comparing the following features: -Just RGB histogram (blue balls); Just greycolor-normalized RGB histogram (pink crosses); Greycolor-normalized RGB histogram divided in two by the waist of each person detection (red circles)

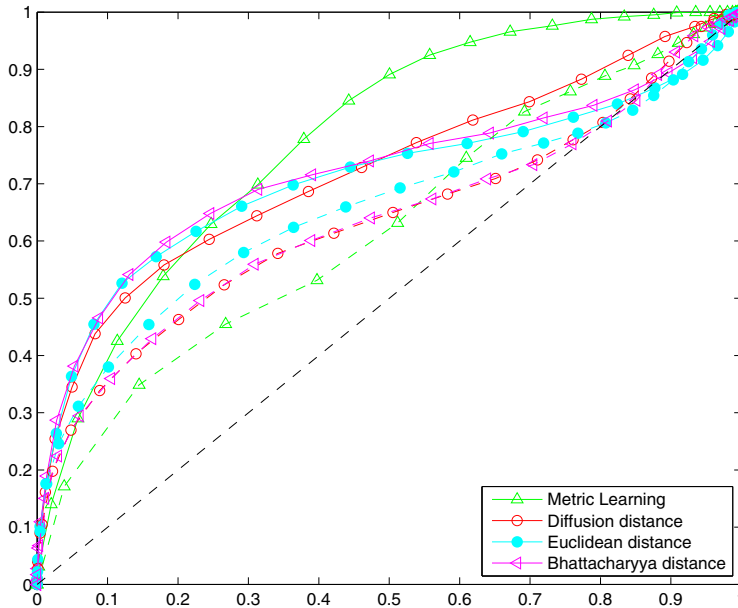


Fig. 4. ROC curves for comparing the different techniques, and confirm the improvement in the results of our suggested feature. Full line: using our waist division feature; Dashed line: not using waist division.

5.3 One to Many Experiments

In Figure 4 we plot the ROC curves for the problem of binary classification “same/not-same pair?” described in Sub-Section 2.2. Initially, distances from all tracks to all tracks are computed. Then a threshold, that determines “similarity” is varied. For each threshold value a True Positive Rate value and a False Positive Rate is computed. All these values plot the Receiver Operating Characteristic (ROC) curves shown in the following Figure 4.

We see in Figure 4 that the baseline Euclidean distance rivals diffusion distance or bests the other techniques. Metric learning performs better as expected since it uses additional information. Learning does seem to improve re-identification and despite requiring labeled data for metric learning, this needs only be done once per system configuration.

Comparing the use of our waist-division feature (full-line) with the counterpart of a single histogram per detection (dashed-line) it is clear the positive influence our feature has in the results of all techniques.

5.4 One to One Experiments

In Figure 5 we plot the Cumulative Matching Characteristic curve for all the techniques implemented, and also analyze the effect of our suggested improvement on the feature, and confirm the improvement in the results of our suggested feature.

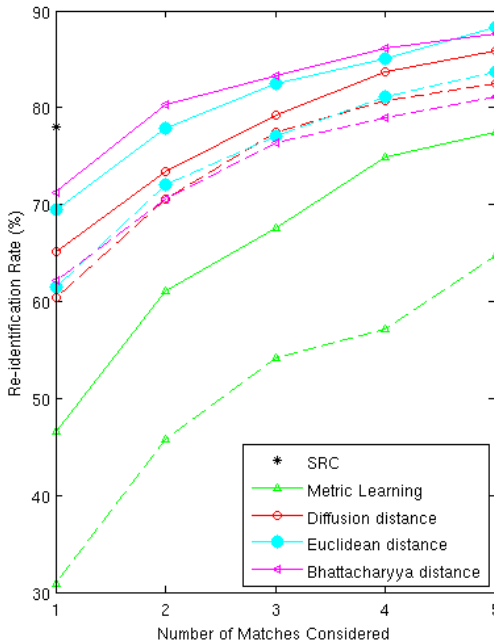


Fig. 5. Cumulative Matching Characteristic curves for comparing the different techniques, and confirm the improvement in the results of our suggested feature. Full line: using our waist division feature; Dashed line: not using waist division.

the waist division described in Sub-Section 3. Using our feature improves the results on all techniques.

Due to the nature of the SRC algorithm, that outputs a sparse solution of one training class per test sample, it is not possible to plot a Cumulative Matching Characteristic curve. Nevertheless SRC gives the best results on the nearest-neighbor level.

6 Conclusions

In this work we addressed the re-identification problem. We not only consider the classical one to one problem, but also the one to many case that may be of interest for practical surveillance applications. In this case someone tries to find similar matches in the system of a given person's image; and the one-to-one approach, where given a person detection we recognize it in a database.

We built upon previous work [10], enriching the feature used by considering upper and lower body parts separately, thus improving the re-identification results.

Metric learning showed promising results in the one-to-many approach, expectedly better than the other distances since it makes use of more information (labeled similarity/dissimilarity pairs). In the one-to-one approach SRC reports the best re-identification rates.

Simple Euclidean distance rivaled or bested the other techniques for re-identification in both approaches.

6.1 Future Work

In the future, we will further enrich the set of features by either adding further body-part selection or integrating SIFT or SURF features to form a multi-modal feature. We will also take advantage of spatiotemporal constraints and appearance correlations between cameras, to limit the search space of re-identification, reducing errors.

We make available by request the matlab source code of the methods developed in this paper as well as the image database produced.

Acknowledgements

This work was supported by project the FCT (ISR/IST plurianual funding) through the PIDDAC Program funds, partially funded with grant SFRH/BD/48526/2008, from Fundao para a Ciênciã e a Tecnologia, and by the project CMU-PT/SIA/0023/2009 under the Carnegie Mellon-Portugal Program.

References

1. Bay, H., Tuytelaars, T., Gool, L.V.: Surf: Speeded up robust features. In: Leonardis, A., Bischof, H., Pinz, A. (eds.) ECCV 2006. LNCS, vol. 3951, pp. 404–417. Springer, Heidelberg (2006)
2. Boulton, T.E., Micheals, R.J., Gao, X., Eckmann, M.: Into the woods: Visual surveillance of noncooperative and camouflaged targets in complex outdoor settings. Proceedings of The IEEE 89, 1382–1402 (2001)

3. Comaniciu, D., Ramesh, V., Meer, P.: Real-time tracking of non-rigid objects using mean shift. In: Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2000, vol. 2, pp. 142–149 (2000)
4. Donoho, D.L.: For most large underdetermined systems of linear equations the minimal ℓ_1 -norm solution is also the sparsest solution. *Comm. Pure Appl. Math.* 59, 797–829 (2004)
5. Hamdoun, O., Moutarde, F., Stanculescu, B., Steux, B.: Person re-identification in multi-camera system by signature based on interest point descriptors collected on short video sequences, pp. 1–6 (September 2008)
6. Javed, O., Rasheed, Z., Shafique, K., Shah, M.: Tracking across multiple cameras with disjoint views. In: Proceedings of Ninth IEEE International Conference on Computer Vision, 2003, vol. 2, pp. 952–957 (2003)
7. Ling, H., Okada, K.: Diffusion distance for histogram comparison. In: CVPR 2006: Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 246–253. IEEE Computer Society, Washington, DC, USA (2006)
8. Lowe, D.G.: Distinctive image features from scale-invariant keypoints (2003)
9. Teixeira, L.F., Corte-Real, L.: Video object matching across multiple independent views using local descriptors and adaptive learning. *Pattern Recognition Letters* 30(2), 157 (2009); video-based Object and Event Analysis
10. Truong Cong, D.N., Achard, C., Khoudour, L., Douadi, L.: Video sequences association for people re-identification across multiple non-overlapping cameras. In: Foggia, P., Sansone, C., Vento, M. (eds.) ICIAP 2009. LNCS, vol. 5716, pp. 179–189. Springer, Heidelberg (2009)
11. Wright, J., Yang, A., Ganesh, A., Sastry, S., Ma, Y.: Robust face recognition via sparse representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 31(2), 210–227 (2009)
12. Xing, E.P., Ng, A.Y., Jordan, M.I., Russell, S.: Distance metric learning, with application to clustering with side-information. In: Advances in Neural Information Processing Systems, vol. 15, pp. 505–512. MIT Press, Cambridge (2002)

Statistical Significance Based Graph Cut Segmentation for Shrinking Bias

Sema Candemir and Yusuf Sinan Akgul

Gebze Institute of Technology
Department of Computer Engineering
GIT Computer Vision Lab., Kocaeli, Turkey
{scandemir,akgul}@bilmuh.gyte.edu.tr
<http://vision.gyte.edu.tr/>

Abstract. Graph cut algorithms are very popular in image segmentation approaches. However, the detailed parts of the foreground are not segmented well in graph cut minimization. There are basically two reasons of inadequate segmentations: (i) Data - smoothness relationship of graph energy. (ii) Shrinking bias which is the bias towards shorter paths. This paper improves the foreground segmentation by integrating the statistical significance measure into the graph energy minimization. Significance measure changes the relative importance of graph edge weights for each pixel. Especially at the boundary parts, the data weights take more significance than the smoothness weights. Since the energy minimization approach takes into account the significance measure, the minimization algorithm produces better segmentations at the boundary regions. Experimental results show that the statistical significance measure makes the graph cut algorithm less prone to bias towards shorter paths and better at boundary segmentation.

Keywords: Graph Cut Segmentation, Energy Minimization, Shrinking Bias, Statistical Significance Analysis.

1 Introduction

Current state-of-the-art segmentation methods are based on optimization procedure [1]. One of the popular optimization based methods is graph cut minimization [2]. The graph cut approach models the image segmentation problem as pixel labeling such that each pixel is assigned to a label which denotes the segmentation classes. The algorithm first builds a graph $G = (V, E)$. V consists of set of vertices that correspond to the pixel features (e.g. intensity) and two extra vertices which denote object and background terminals. E consists of edges which are assigned to a nonnegative weights according to the relationship between the vertices. After the graph structure is constituted, the optimal labeling configuration is found by minimizing an energy functional whose terms are

based on the edge weights of the graph. The standard graph energy functional is formulated as,

$$E(f) = \sum_{i \in V} E_d(f_i, d_i) + \lambda \sum_{i, j \in N} E_s(f_i, f_j), \tag{1}$$

where V are the vertices, f_i is the segmentation label, d_i is the a priori data of pixel i , and N represents the neighborhood pixels j of pixel i . The first term in the energy functional is called the data term E_d , which confines the segmentation labels to be close to the observed image. The second term is used for the smoothness which confines the neighboring nodes to have similar segmentation labels. The regularization weight λ balances the relationship between the data and smoothness terms.

1.1 Motivation

Graph cut algorithms produce successful solutions for the image segmentation [2,3,4]. However, the foreground boundary, especially at the detailed parts still cannot be obtained well in the graph cut minimization. There are basically two reasons of inadequate segmentations at the boundary regions:

(i) Data-Smoothness Relationship. One of the reasons of the inadequate segmentation of graph cut algorithms is due to the energy minimization approach. The trade off between the data and the smoothness terms should be well regularized in the energy functional. In order to obtain the boundary of the foreground accurately, regularization should be small. In Fig 1.b, segmentation is obtained with a small λ . Small λ segments the objects sharply, however, it produces noisy solutions (grassy regions). If we increase the λ in order to obtain a noiseless segmentation, this time we lose the details such as the legs and the ears of the horses (Fig 1.c). For the optimal segmentation, λ parameter should be optimal as in Fig 1.d. Even for the optimal segmentation, the detailed parts of the foreground still cannot be segmented accurately. The main reason of the inadequate segmentation in energy minimization approach is that the optimal regularization parameter for overall segmentation is generally high for the boundary regions.

(ii) Shrinking Bias. Another reason of the inadequate segmentation of graph cut minimization is the shrinking bias [5] which is an inherent bias towards shorter paths. The smoothness term in graph-cut methods consists of a cost summation over the boundary of the segmented regions. A short expensive boundary may cost less than a very long cheap one. Especially at the long and thin boundaries of objects, the graph cut algorithms may cut the boundary along the shorter paths which causes inadequate segmentation for those parts. Figure 2 shows the optimal segmentation for the horse image in Figure 1.a and illustrates the shrinking bias problem. The green boundary denotes the ground truth segmentation. However, the graph cut algorithm segments the image along the red boundary. Note the marked regions on the image. The algorithm segments the object at the short-cut boundaries instead of long and thin boundary paths.



Fig. 1. The illustration of the trade off between the data and the smoothness terms of the graph cut minimization. a) Input image. b) Segmentation by a small λ . Less regularization provides to segment the detailed regions of the foreground such as the legs parts. c) Segmentation by a large λ . Not only the noisy segmentation but also the detailed parts of the segmentation is lost. d) Optimal segmentation is still not well enough at the boundary parts.

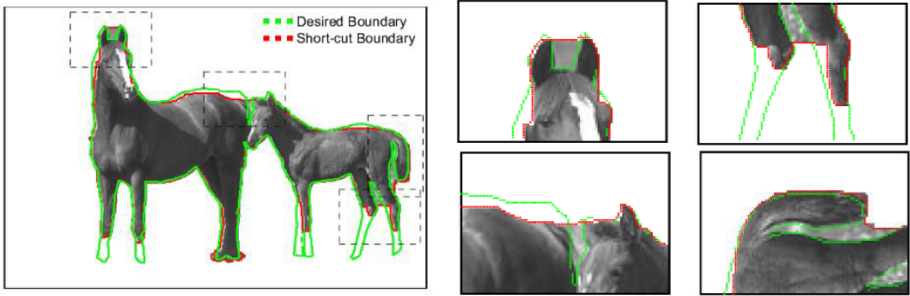


Fig. 2. Graph Cut methods may short-cut the foreground along the red borders instead of following the green borders, because short-expensive boundary may cost less than a very long cheap one

1.2 Related Work

Shrinking bias problem of graph cuts is first addressed by Kolmogorov and Boykov [5]. They define the flux along the boundary and improve the segmentation. Flux knowledge causes stretching at the boundary while the graph cut algorithm tries to smooth the solution because of the energy minimization. Although the flux integration produces better solutions than the original graph cut approach, the algorithm cannot be extended to color images, because flux can be defined only on the grey-level images [6]. Another work which tries to overcome the inadequate segmentation is geodesic segmentation which avoids the shrinking bias of the graph cut methods by removing the edge component in the energy formulation [7]. However this approach cannot localize the object boundaries and it is very sensitive to seed placement [8]. Vincente and Kolmogorov [6] attempt to solve the long and thin object segmentation by adding connectivity priors. They manually add some additional marks at the endpoints of long-thin objects and then run the Dijkstra's algorithms after the graph cut minimization. Recently, researchers argued that the same λ may not be optimal for all regions of the

image. They proposed different algorithms which spatially change the regularization parameter based on the local attributes of the images [9,10,11,12]. Since the regularization weight is decreased at the boundary parts, energy minimization cannot over-smooth the thin and long parts of the foreground. Therefore, the spatially-adaptive methods will produce better segmentation results than the traditional graph cut algorithms.

In this work, the statistical significance measure is integrated into the energy minimization approach in order to improve the image segmentation problem. In traditional statistics, statistical significance measures the randomness of an outcome. It is previously proposed that the statistical significance can be used as a comparison measure for the outcomes of different distributions [13,14]. In this work, we redefine and modify the idea for the shrinking bias problem and include additional experiments. The statistical significance measure is included in the energy minimization approach through the graph structure. We measure the statistical significance of all weights on the graph. Then we reconstruct the graph structure by changing the weights with their statistical significance measurements.

2 Statistically Significant Graph Cut Segmentation

2.1 *p*-value Calculation

Statistical significance is a probability value (*p*-value) which is the measurement of randomness. It is used for the hypothesis testing mechanism in statistics. If the observed outcome of an experiment is statistically significant, this means that it is unlikely to have occurred by chance, according to the significance level which is a predetermined threshold probability.

In order to measure the statistical significance of the outcome of an experiment, cumulative probability distribution function of the experiment should be known. If the distribution of the outcome is a known distribution such as the exponential distribution, the parameters of this distribution is used to measure the significance. On the other hand, if the distribution is not known, the possible outputs of the experiment is used to form the probability distribution. The area under the probability distribution forms the cumulative distribution function. The location of the outcome on cumulative distribution determines the statistical significance of the observed outcome. Equation 2 denotes the statistical significance of the outcome *x*.

$$F(x) = P(X \leq x) = \sum_{-\infty}^x P(X = x) \tag{2}$$

$P(X = x)$ is the probability distribution of experiment *X*, $F(x)$ produces the *p*-value of the statistic *x*. If the obtained *p*-value is small then it can be said that an unusual outcome has been obtained.

2.2 Measuring the Significance of Edge Weights

In this work, we used the significance measure to bring the data and smoothness energy terms (outcomes) into the same base, which is different from the traditional usage. We measured the statistical significance of energy terms in terms of edge weights of the graph structure. In graph cut algorithms, objective function is constituted of the edge weights. The edges between the terminal and pixel vertices are called t -links whose weights form the data energy term. On the other hand, the edges between the neighboring pixel vertices are called n -links whose weights form the smoothness terms of the energy function. The weights of different types of links are determined through different functions such as squared differences, absolute differences, truncated absolute differences, laplacian zero crossing or gradient direction. As an example, in the interactive segmentation of Boykov and Jolly [2], the weights of t -links are based on the marked pixel histogram, whereas, n -links are the intensity difference between the neighboring pixels. Note that, data and smoothness terms of the energy formulation have different functional forms, whereas, graph cut minimization try to minimize the different functional forms simultaneously through the same objective function. In this work, we used the significance measure to bring the energy terms on the common base by expressing the weights in terms of the statistical significance measure.

In order to measure the statistical significance of data and smoothness terms, the probability distribution of the terms should be generated. The edge weights on the graph form the probability distribution of terms. Figure 3 illustrates the procedure. The weights of the t -links (marked as red color on the graph) form the probability distribution of data term of the energy function, on the other hand, the weights of the n -links (marked as blue on the graph) form the probability distribution of smoothness term. Two sample edge weight is denoted on the graph by green color. Then we measure the statistical significance of each edge weight by evaluating the weights on the distributions. t -link weights are evaluated on the data term distribution; n -link weights are evaluated on the smoothness distribution. After measuring each weight significance, we reconstruct a new graph structure in which edge weight is assigned to a significance value.

Equation 3 and Equation 4 formulates the significance measurement.

$$F(x_d) = P(E_d(f, d) \leq x_d), \quad x_d = E(f_i, d_i) \quad (3)$$

$$F(x_s) = P(E_s(f, d) \leq x_s), \quad x_s = E(f_i, f_j) \quad (4)$$

where x_s is the observed data weight, x_s is the observed smoothness weight, $P(E_d(f, d))$ denotes the probability distribution of data weights, and $P(E_s(f, d))$ denotes the probability distribution of smoothness weights.

3 Data-Smoothness Weights Relationship

We measure the statistical significance of each term by evaluating the terms according to the other graph terms. Evaluating the terms on its own distributions

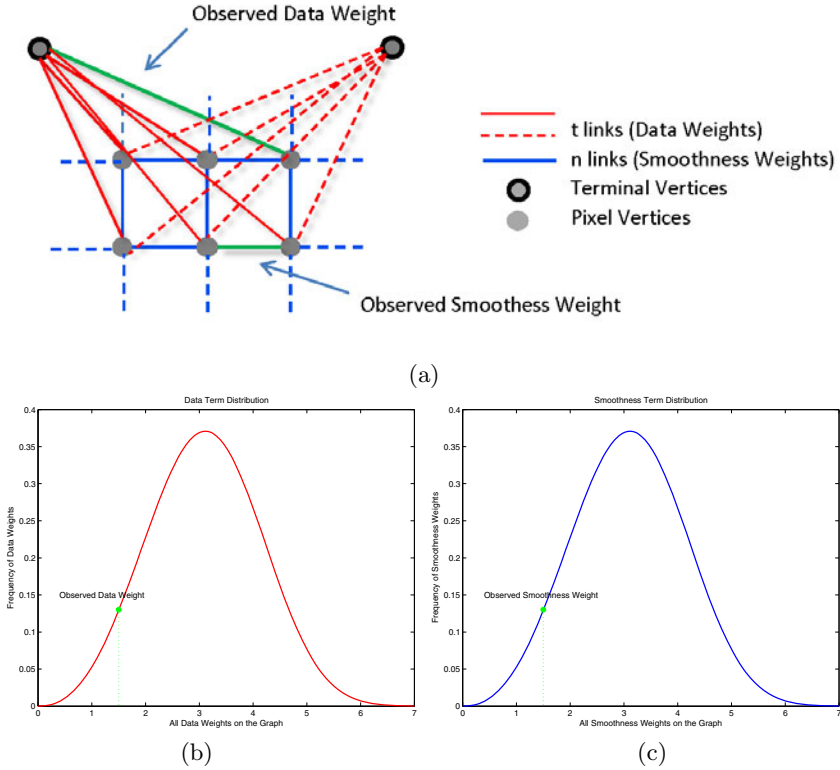


Fig. 3. Data and smoothness weights are normalized by evaluating weights according to other data and smoothness weights on the graph. a) A simple graph structure. Data edges denoted by red color, smoothness edges denoted by blue color. b) Probability distribution of data terms. c) The probability distribution of the smoothness terms.

and expressing the edge weights by the same measurement have two explicit advantages:

(i) The significance measure decreases the scale and distribution differences between the data and smoothness energy terms and bring them on similar base. Therefore, the tradeoff between the terms would be properly regularized.

(ii) The significance measure for the data weights are determined according to other data weights on the graph. Similarly, the significance measure for the smoothness weights are determined according to other smoothness weights on the graph. It can be interpreted as each weight is normalized relative to other weights. As an example, if one of the data weight has a high significance among the other data weights, we can say that data term for that pixel is statistically more significant than the smoothness term, albeit both terms have equal weight. Normalization change the relative weights of data and smoothness terms according to their randomness. Rare weights become more important than the normal weights.

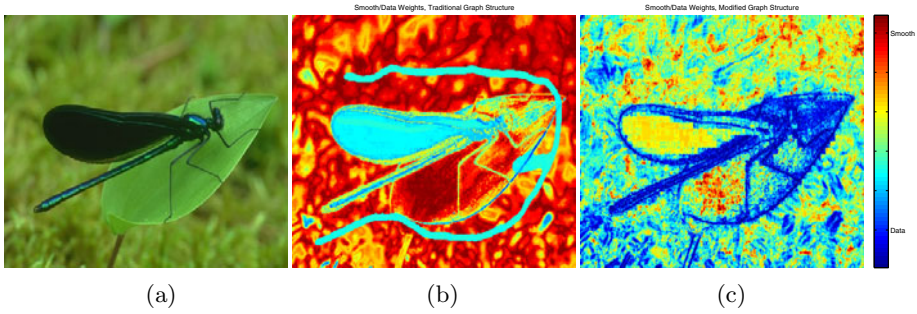


Fig. 4. Data-Smoothness relationship for the dragonfly image. (a) A sample image. (b)Smoothness/Data weight rate for each pixel of traditional graph structure. (c) Smoothness/Data weight rate for each pixel of modified graph structure.

In order to show the relative relationship between data and smoothness weights of each pixel, we constructed weight maps for both graph structures. We calculated the relative weight of each pixel $i \in I$ of image I using Formula 5. Then we normalize the weight rates to a fixed range [0-1]. If the weight rate is close to the 1, this means that smoothness weight is relatively bigger than the data weight for that pixel. If the smoothness weight increases, pixel get closer to the red. On the hand, if the weight rate is close to the 0, it can be said that the data weight is more important for that pixels. We show that type of pixels with blue. Figure 4a denotes the weight map of original graph structure, Figure 4b denotes the weight map of modified graph structure. Note that data weights at the boundary part takes more importance than the smoothness weights in the modified graph structure.

$$\frac{\lambda E_s(f_i, f_j)}{E_d(f_i, d_i)} \quad \forall i \in I \tag{5}$$

4 Improvement in Shrinking Bias Problem

Statistical significance measurement decreases the smoothness weights along the boundary as it can be seen in Figure 4. Therefore finding a short expensive boundary, which may cost less than a very long cheap one become harder. Figure 5 demonstrates the improvement in shortcutting. The segmentation is obtained by minimizing the modified graph cut structure whose weights are calculated by significance measurement. The red contour denotes the short-cut boundary which is the optimal segmentation of traditional graph structure as we showed previously in Figure 2. Note that the blue contour is explicitly closer to the desired boundary.

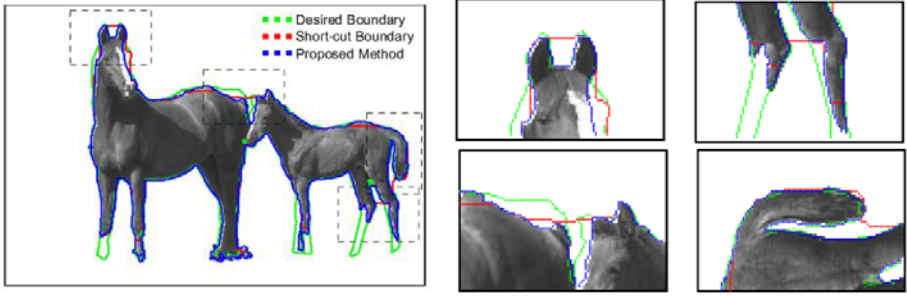


Fig. 5. Improvement in Shrinking Bias problem. The blue contour is closer to the desired boundary than the short-cut boundary.

5 Experimental Results

To quantitatively evaluate the accuracy of the segmentation of the proposed approach, we applied it to the Berkeley dataset [15]. We obtained optimal segmentation of original graph structure and modified graph structure by comparing the segmentations by ground truths. Figure 6 displays some segmentations results of both approaches. Note that the thin and long parts of foreground such as legs of the dragonfly or wood in the bear image. The proposed approach produces better solutions at these problematic parts. The percentage errors of the segmentations are listed on Table 1.

Table 1. Error Rates of the Segmentations in Figure 6

Image	Traditional Graph Structure Optimal Segmentation Error Rate	Modified Graph Structure Optimal Segmentation Error Rate
Dragonfly	1.29 %	1.08 %
Eagle	4.34 %	2.99 %
Horse	2.97 %	2.56 %
Bear	4.32 %	3.07 %
Plane	0.83 %	0.70%
Trees	1.74 %	1.01%

6 Discussion

In this paper we have integrated the statistical significance measure into the graph structure in order to improve the graph cut segmentation approach. We measured the significance of data and smoothness edge weights according to other weights. Then we constructed a new graph structure whose edge weights are the significance measurements. Using the significance measurements instead of weights can be interpreted as each weight is normalized relative to other weights. In the new

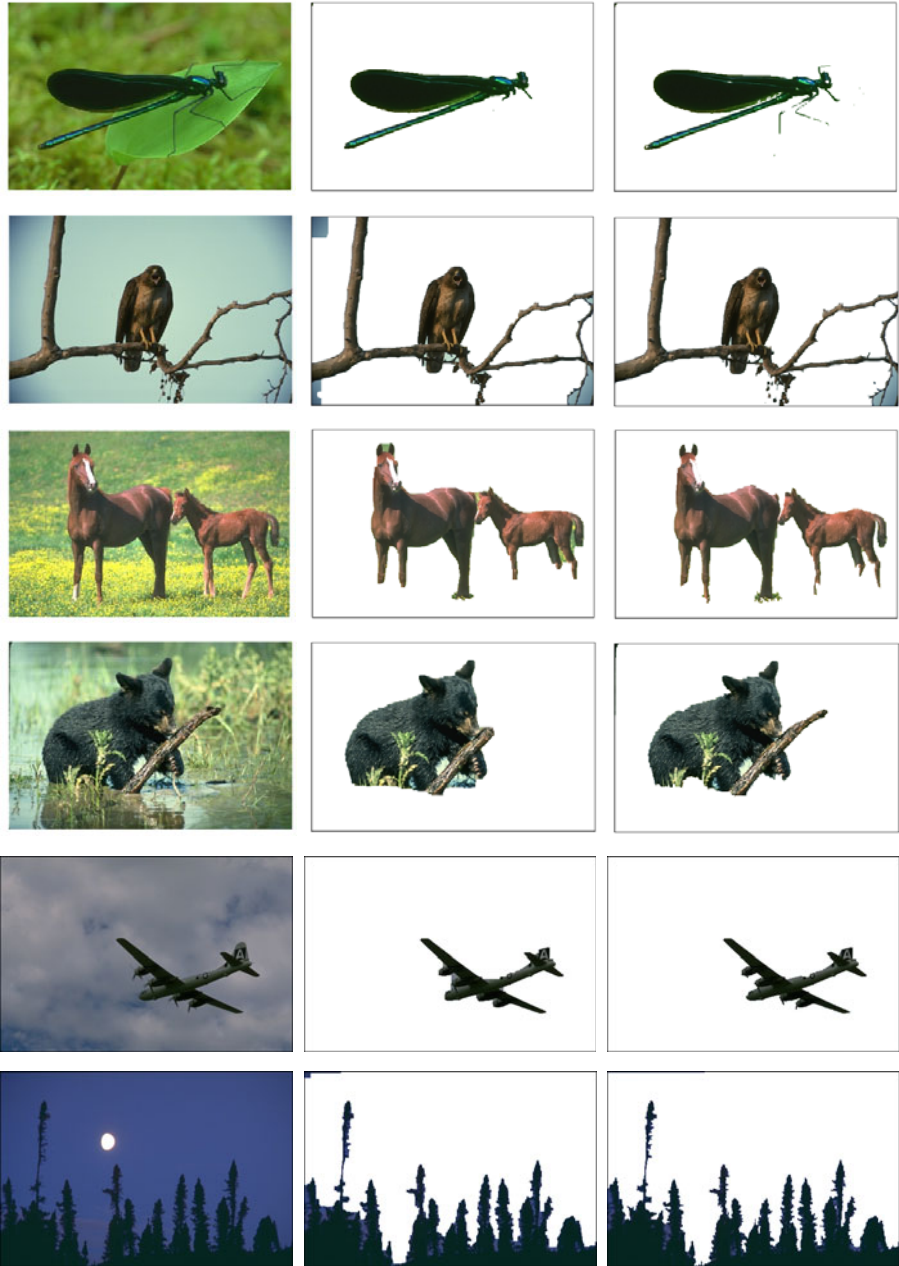


Fig. 6. FirstColumn: Image from the Berkeley set [15]. SecondColumn: Optimal segmentation by traditional graph structure. ThirdColumn: Optimal segmentation by modified graph structure based on statistical significance measurement.

graph structure, the relative weights of data and smoothness edges are changed according to their randomness. Especially at the boundary regions of the foreground, the data weights gets more importance than the smoothness weights. In another word, the smoothness weights along the boundary is decreased. Therefore, finding a short expensive boundary which may cost less than a very long cheap one become harder. We demonstrated our algorithm on several images on Berkeley segmentation set, and showed that our optimal segmentations are better than the optimal segmentations of traditional graph cuts.

References

1. Felzenszwalb, P., Zabih, R.: Dynamic Programming and Graph Algorithms in Computer Vision. To appear in the IEEE Transactions on Pattern Analysis and Machine Intelligence
2. Boykov, Y.Y., Jolly, M.-P.: Interactive Graph Cuts for Optimal Boundary and Region Segmentation of Objects in N-D Images. In: ICCV (2001)
3. Boykov, Y., Funka-Lea, G.: Graph cuts and efficient N-D image segmentation. *Int. J. Computer Vision* 70, 109–131 (2006)
4. Rother, C., Kolmogorov, V., Blake, A.: Grabcut: Interactive foreground extraction using iterated graph cuts. In: SIGGRAPH (2004)
5. Kolmogorov, V., Boykov, Y.: What metrics can be approximated by geo-cuts, or global optimization of length/area and flux. In: IEEE ICCV, vol. 1, pp. 564–571 (2005)
6. Vicente, S., Kolmogorov, V., Rother, C.: Graph cut based image segmentation with connectivity priors. In: IEEE CVPR (2008)
7. Bai, X., Sapiro, G.: A geodesic framework for fast interactive image and video segmentation and matting. In: IEEE ICCV, pp. 1–8 (2007)
8. Price, B.L., Morse, B., Cohen, S.: Geodesic Graph Cut for Interactive Image Segmentation. In: IEEE CVPR (2010)
9. Rao, J., Hamarneh, G., Abugharbieh, R.: Adaptive Contextual Energy Parameterization for Automated Image Segmentation. In: Bebis, G., Boyle, R., Parvin, B., Koracin, D., Kuno, Y., Wang, J., Wang, J.-X., Wang, J., Pajarola, R., Lindstrom, P., Hinkenjann, A., Encarnação, M.L., Silva, C.T., Coming, D. (eds.) ISVC 2009. LNCS, vol. 5875, pp. 1089–1100. Springer, Heidelberg (2009)
10. Gilboa, G., Darbon, J., Osher, S., Chan, T.: Nonlocal convex functionals for image regularization, UCLA CAM-report (2006)
11. Candemir, S., Akgül, Y.S.: Adaptive Regularization Parameter for Graph Cut Segmentation. In: Campilho, A., Kamel, M. (eds.) ICIAR 2010. LNCS, vol. 6111, pp. 117–126. Springer, Heidelberg (2010)
12. Schoenemann, T., Kahl, F., Cremers, D.: Curvature regularity for region-based image segmentation and inpainting: A linear programming relaxation. In: IEEE ICCV (2009)
13. Candemir, S., Akgul, Y.S.: A Nonparametric Statistical Approach for Stereo Correspondence. In: IEEE Conf. Computer and Information Sciences, pp. 1–6 (2007)
14. Candemir, S., Akgul, Y.S.: Statistical Significance Based Graph Cut Regularization for Medical Image Segmentation. To appear in Turkish Journal of Electrical Engineering and Computer Sciences
15. The Berkeley Segmentation Dataset and Benchmark Page, <http://www.eecs.berkeley.edu/Research/Projects/CS/vision/grouping/segbench/>

Real-Time People Detection in Videos Using Geometrical Features and Adaptive Boosting

Pablo Julian Pedrocca and Mohand Saïd Allili

Université du Québec en Outaouais,
Département d'Informatique et d'Ingénierie,
101, Rue St-Jean-Bosco, Gatineau, QC, J8X 3X7, Canada
{pedp01,mohandsaid.allili}@uqo.ca

Abstract. In this paper, we propose a new approach for detecting people in video sequences based on geometrical features and AdaBoost learning. Unlike its predecessors, our approach uses features calculated directly from silhouettes produced by change detection algorithms. Moreover, feature analysis is done part by part for each silhouette, making our approach efficiently applicable for partially-occluded pedestrians and groups of people detection. Experiments on real-world videos showed us the performance of the proposed approach for real-time pedestrian detection.

Keywords: People detection, geometrical features, AdaBoost.

1 Introduction

Automatic human detection is a key issue for many computer vision applications, such as robotics, video surveillance, human computer interaction and automated person assistance [1]. Recently, several approaches have been proposed for people modeling and detection in videos. These approaches can be broadly classified into two main groups.

In the first group, methods based on histograms of oriented gradients (HoG) are used to learn the shape of humans using techniques such as Adaboost [2-5] or support vector machines (SVM) [6, 7]. The main advantage of those methods is that the detection relies on features based on the derivatives of the image, which are less sensitive to variations of human appearance. These features put together in high-dimensional vectors are assumed to capture the pedestrian shape. However, if an object is partially occluded, or the pedestrian walks against a cluttered background, or in a group of people, the detection may fail since a great part of the pedestrian boundary will be missed. In the second group, learning techniques are used with features similar to Haar wavelets [8, 9]. The success of those methods relies on the assumption that the pedestrians walk against a uniform background, allowing block differences (i.e., wavelets-like features) to capture the different parts of a pedestrian. Therefore, the contour must be strong enough to distinguish the pedestrian from its immediate neighborhood in the image.

Besides, since the detection relies on block differences, the variation in pedestrian versus background appearance will influence the description, which is not desirable for a detection that is robust to appearance changes, like the methods in the first group.

One major limitation in the above methods lies in the fact that they are dedicated to single, non-occluded, person detection. If several persons walk in a group, or in the occurrence of occlusions, the detection can fail since a great part of the boundary of the person will be lost (in the group). For realistic scenarios, it is important that a pedestrian detection algorithm should work for both isolated humans and groups of people. Last but not least, the above methods require to scan the whole image and test for every possible window if it contains a pedestrian. This can be computationally expensive and, therefore, not suitable for real-time applications like video surveillance.

In the present paper, we propose an efficient and real-time approach for people detection in videos. Based on the output of change detection algorithms, we use directly the generated silhouettes to detect if a moving object is a pedestrian or not. We rely for our detection on using example-based learning approach where a model of pedestrians is generated. First, each silhouette is analyzed geometrically, and part by part, to detect if it contains possible humans. In other words, we detect if a blob corresponds to a group of people, to a person with another object (e.g., a person on a bicycle, etc.), or if it does not contain any human. If a blob is likely containing humans, a second analysis is performed on the blob in order to isolate each person. For each step, we use AdaBoost algorithm to learn a model from several examples of human silhouettes, and use the model to classify each new silhouette.

This paper is organized as follows: Section 2 presents the main steps composing our approach. Section 3 presents some experiments on real world videos. We end the paper with a conclusion and future work perspectives.

2 The Proposed Approach

In what follows, we present the different modules that compose our algorithm. Starting from the first frame of the sequence, an adaptive background model is built for the sequence as time goes on. Given a new frame of the sequence, a change detection is operated first to separate the moving objects from the rest of the image. Then, a first analysis is performed on the resulting blobs in order to detect the ones that may contain people. Finally, the analysis is refined on those blobs in order to isolate each person. Fig. 1 gives an overview of the different modules composing our algorithm. In the following sections, we develop separately the details of each module.

2.1 Background Subtraction

Since our approach uses object silhouettes, a good background subtraction algorithm is essential to the success of our detection. In the literature, there are several change detection algorithms that vary in complexity and efficiency. They

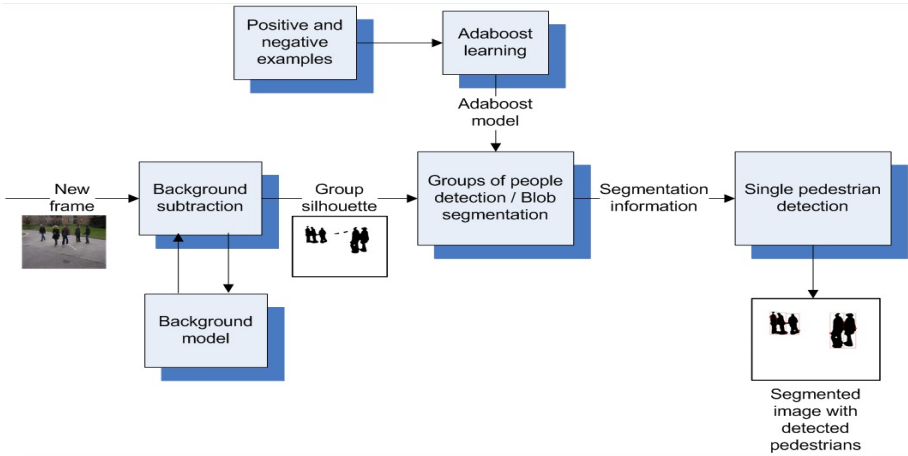


Fig. 1. Architecture for our moving pedestrian detection in videos

range from the most naïve absolute difference between pixels to more complicated algorithms that create a statistical model of the image, and process individual pixels based on their historic rate of change [10–13]. While the former is fast, but lacks efficiency, the latter is expensive in computation time and memory. To achieve a fast, yet efficient, change detection we propose an approach that focuses on the regions with the highest occupancy, aiming to process only the regions that interest us (i.e., where the presumed pedestrians are likely to be located). We achieve that by concentrating in the areas of the image where the change density is higher, i.e., where a large number of pixels has changed. This aims to exclude regions with relatively small changes due to non-stationary backgrounds (i.e., swaying trees, etc.). Therefore, it allows to filter most noise and image artifacts while focusing on potential humans.

First, we initialize the background model \mathbf{B} with the first image of a given video sequence. In ideal conditions, this image will not contain moving objects. The algorithm then starts to process all subsequent images. For each image \mathbf{I} , our algorithm calculates a coarse foreground mask \mathbf{M} and its complement \mathbf{M}' . The process to calculate \mathbf{M} is the following: 1) The image \mathbf{I} is divided by the background model \mathbf{B} and multiplied by a fixed coefficient C . 2) The resultant matrix is then filtered by hysteresis [14]. When filtered by hysteresis, an upper threshold and a lower threshold are applied. In our example, we apply a range $C \pm R$, where $C+R$ is the upper threshold and $C-R$ is the lower threshold. Both C and R are values that are determined empirically (please see the experimental results section). 3) The resultant mask is dilated using a morphological filter, where holes are filled as well. This preliminary coarse subtraction provides the regions that are most different between the background model \mathbf{B} and the current image \mathbf{I} . We call this mask \mathbf{M} . Then, the absolute difference between \mathbf{I} and \mathbf{B} is calculated for the pixels in \mathbf{M} and then thresholded.

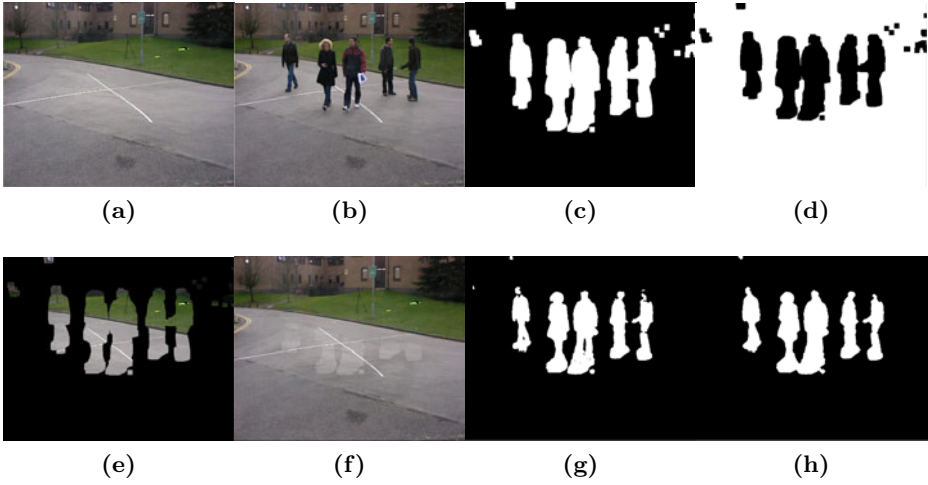


Fig. 2. a) Original background model \mathbf{B} ; b) Frame to be processed \mathbf{I} ; c) Coarse mask \mathbf{M} obtained after the first filtering; d) Mask complement \mathbf{M}' ; e) Image resulting of multiplying the background model by the mask $\mathbf{B} \times \mathbf{M} = \mathbf{I}'$; f) Reconstructed background model \mathbf{B}_{t+1} ; g) Subtracted image; h) Final blobs after morphological dilation and hole filling

Fig. 2 shows an example of background subtraction using our approach. To update the background model, the algorithm takes the current image \mathbf{I} and multiplies it by the mask complement \mathbf{M}' . This provides the current background \mathbf{I}' without the moving blobs. The background model \mathbf{B}_t is multiplied then by the mask \mathbf{M} , and that provides the background model \mathbf{B}' that belongs only to the blob section. The combination $\mathbf{I}' + \mathbf{B}'$ gives the updated background model \mathbf{B}_{t+1} .

We compared different background methods against our approach to find the most suitable one. This comparison (by no means exhaustive) considered the frame difference (FD) method, the approximate median (AM) method, and the mixture of Gaussians (MoG) method. The major flaw of all those methods against ours is that, even though the background model is dynamically updated, objects tend to fade out and become part of the background as time goes along. Fig. 3 shows example illustrating this fact. We can note that the chair disappears after a number of frames (see the second column of the figure). The frame difference method generates a trailing that increases the effective surface of the silhouette. MoG provides, in general, better results than FD and AM, but it is computationally very expensive. Among the compared methods, our approach provides the most sharp and regular silhouette contours, which is very critical for calculating accurate geometrical features as will be explained in the next sections.

2.2 Groups of People Detection

Once the candidate blobs have been identified, a blob segmentation process is necessary to determine whether a single person or a group of persons are contained within each blob. One approach used by [15] is to obtain a projection histogram of the blob. The projection histogram calculates the sum of all positive pixels on a given row. Unfortunately, such approach is not always good, for example, when the pedestrian is leaning to one side and the head is not in line with the rest of the body.

The approach we took is to trace a blob curve based on the distances of the first positive pixel (on a vertical direction) to the horizontal line passing through the center of the blob. The lowest points in the blob will give the lower curve values, and vice versa. The curve will then be equal as the blob's top boundary profile (see Fig. 4). Once the curve is established, we detect its peaks

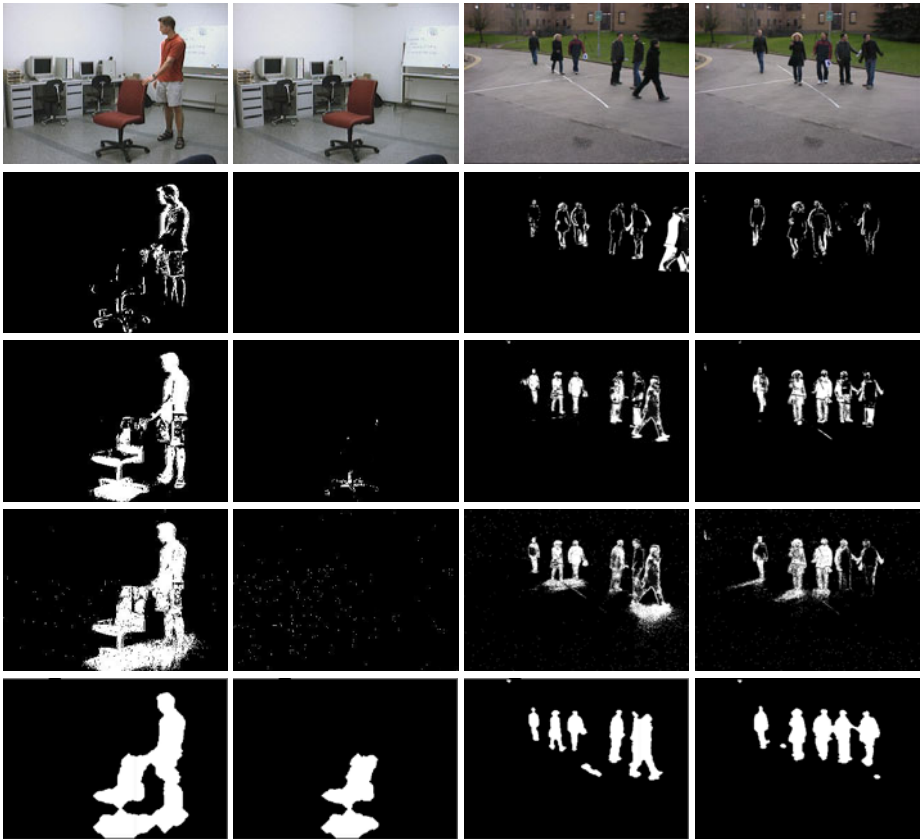


Fig. 3. Comparison between different background subtraction methods. The first row shows the original images, whereas the rows below show, respectively, the frame difference method, the approximate median method, the mixture of Gaussians method, and our proposed approach.

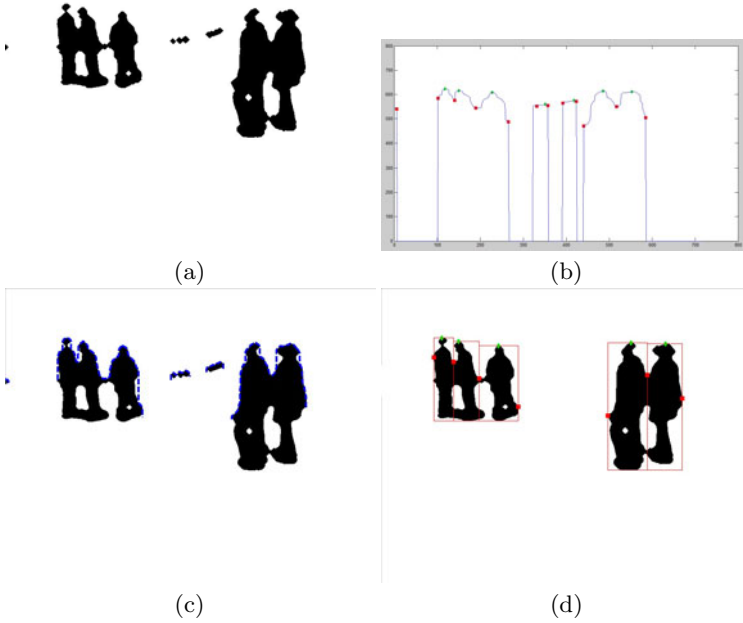


Fig. 4. a) Source image; b) Horizontal profile with maximum and minimums already marked; c) Image with profile line superimposed (only over the detected blobs); d) Segmented blobs

and valleys. Then, we segment each blob using the peaks as the middle of the pedestrian candidate and the peak's adjacent valleys as the pedestrian lateral limits. Fig. 4 shows an example of group of people detection. Each pedestrian candidate is surrounded by a red square (see Fig. 4d), the horizontal limits of the squares are given by the blob's valleys, and the presumed pedestrian heads are given by the blob's peaks.

2.3 Geometrical Feature Extraction

The Histogram of Oriented Gradients (HoG) is a method that has been widely used to represent the shape information of the objects. In our work, we use this information for identifying blobs that contain humans. Since we have the silhouettes of moving objects, we extract the HoG only for the supposed pedestrian's head's silhouette that we detect in the blobs. The algorithm takes a segmented blob and the boundary of its upper portion (1/5) that will be the region of interest (ROI) to be used to calculate the HoG. We then build the gradient orientation histogram by grouping the occurrences of each angle using 9 bins. To make the histograms invariant to scale change, we normalize the histograms by dividing each bin by the number of pixels on the object contour.

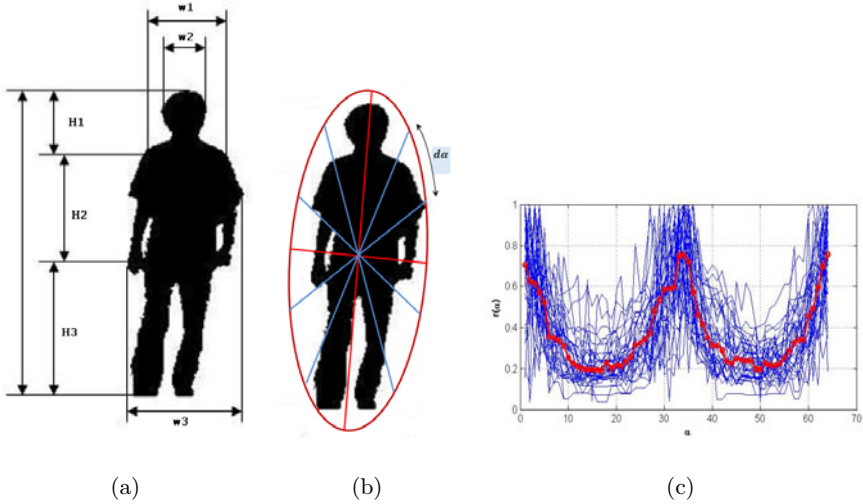


Fig. 5. Geometrical features calculated for each silhouette: a) inter-parts distances, b) global shape signature, c) the signature in b) shown for 200 examples of human silhouettes (the red line represents the median signature)

When a blob likely contain humans, a second analysis will consist of isolating each person contained in the blob. For this purpose, we use geometrical features calculated from each silhouette, where we first segment each blob using the peaks as the middle of the pedestrian candidate and the peak's adjacent valleys as the pedestrian lateral limits, as explains in section (2.2). We then train an AdaBoost model for human silhouettes according to the features $\frac{w_1}{w_2}$, $\frac{w_2}{w_3}$, $\frac{H_1}{H_2}$, $\frac{H_1}{H_3}$, $\frac{H_2}{H_3}$, $\frac{w_3}{H_1+H_2+H_3}$, $\frac{w_2}{H_1}$, $\frac{w_3}{H_3}$ calculated on the blob segments (see Fig. 5.a), and the normalized distance-versus-angle signature $\mathbf{r}(\alpha)$ (see Figs. 5.b and 5.c). For training, we used 400 examples of human silhouettes extracted from real videos.

Since we require a binary classification for our silhouettes, Adaboost algorithm [16] is suitable for this task. Adaboost works by combining several weak linear classifiers to construct a strong classifier. A weak classifier is defined to be a classifier which is only slightly correlated with the true classification (it can label examples better than random guessing). The number of weak classifiers used during training impacts on the correctness of the final classifier. Intuitively, we could say that the more weak classifiers we combine, the more correct the final classifier is expected to be. However, we found that there is a 'sweet spot' in terms of number of classifiers, and increasing the number of rounds does not necessarily have a positive impact on performance (see Table I).

3 Experimental Results

For the training set, we use a dataset of 800 examples (400 positive examples - belonging to humans- and 400 negative examples). Out of 400 positive examples,

250 were extracted from the PETS 2009 sequences after applying background subtraction, and the others from other videos. Negative examples were chosen not to contain human silhouettes but instead objects likely to move in front of a camera, such as vehicles or pets. For our experiments, we obtained experimentally the optimal number of classifiers that is suitable for our algorithm. To establish this number, we trained the model using 20, 30, 40, 50, and 60 weak classifiers, respectively. Then, we used the obtained model to classify 100 labeled data that do not belong to the training set. We found that the combination of 40 weak classifiers gives the best results. We define *Precision* as $\frac{TP}{TP+FP}$, *Recall* as $\frac{TP}{TP+FN}$, and *Accuracy* as $\frac{TP+TN}{TP+TN+FP+FN}$; where TP (True positives) is the number of human blobs detected as such, TN (True negatives) is the number of human blobs classified as such; FP (False Positives) is the number of blobs classified as human when they are not, FN (False Negatives) is the number of blobs classified as non human when they're in fact human. Table 1 gives the values of these criteria for establishing the best number of classifiers:

Table 1. Precision, recall and accuracy vs. number of weak classifiers

Number of weak classifiers:	20	30	40	50	60
Precision	0.869	0.903	0.939	0.931	0.932
Recall	0.923	0.931	0.946	0.938	0.937
Accuracy	0.895	0.917	0.944	0.939	0.938

We tested our algorithm using different video sequences. We used experimental values of $C = 0.5$ for the normalization coefficient and $R = 0.1$ for hysteresis thresholding. After the preliminary difference is multiplied by the mask, we apply an empirical threshold of 0.01 for background subtraction. Then, we apply closing and opening morphological operations using a 5×5 diamond structure. Finally, we smooth out the silhouette image using a 3×3 Gaussian filter in order to obtain regular silhouette contours.

One video we used was "RedChair.avi" [17]. This video has 187 frames and shows a man carrying a chair across a room. This gives us the opportunity of identifying the person alone, or partially occluded, or both (e.g., if the same blob contains a person and non-person object). Fig. 6 shows an example from this video. From left to right and top to bottom, we show frames 54, 67, 60 and 75 of the sequence with the detected blobs in each frame. Green rectangles delimit the blobs identified as pedestrians, whereas red rectangles delimit the blobs identified as non-pedestrian. In the first row, we notice that whether there is either backward or forward occlusion, the detection is positive and the whole blob is marked as pedestrian. In the bottom row, even though they are part of the same blob, both objects are properly detected and classified as pedestrian or non-pedestrian. Finally, the last frame shows a non-human object that is left at the scene and correctly classified as non-pedestrian.

Another video was based on the PETS 2009 benchmark data; we used the dataset *S0/Background/View08* to generate the background model and the dataset *S0/CityCenter/Time12-34/View08* for the moving people. This yields

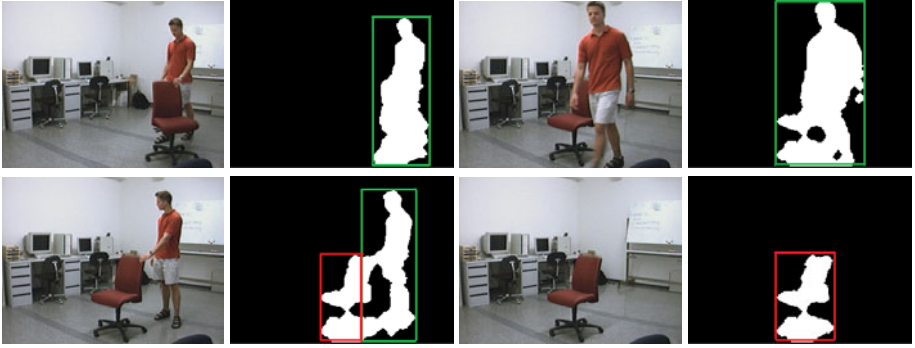


Fig. 6. Example of person detection. We show from left to right and top to bottom frames 54, 67, 60 and 75 of the sequence with the detected blobs for each frame. Green rectangles delimit the blobs identified as pedestrians, whereas red rectangles delimit the blobs identified as non-pedestrian.

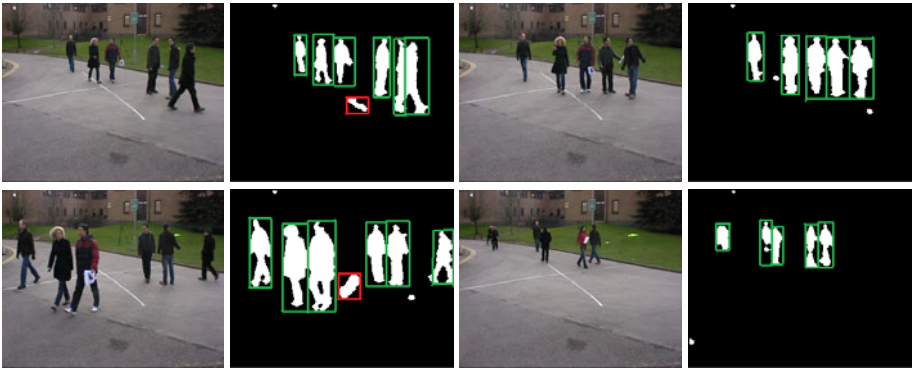


Fig. 7. Example of person/group detection. The left column shows frames 58, 75, 94 and 472 from the PETS 2009 dataset (view 008, time 12:34). The right column shows the detected blobs for each frame.

a sequence with a total of 939 frames which show pedestrians walking either alone, in groups, or meeting to form groups and then dispersing. Fig. 7 shows some examples extracted from this sequence.

Table 2 shows a quantitative performance evaluation of our method compared to the HoG [2] and W4 [15] approaches. We note that the worst method for the precision, recall and accuracy criteria is [15] since it uses critical points for silhouette classification. The most time demanding method is [2] since it scans all the image to detect possible pedestrians. While a comparable performance is obtained for the precision, recall and accuracy criteria, the computational time is much more efficient for our approach than [2]. Consequently, using our method is more advantageous for real-time detection scenarios than using [2, 15]. Our

Table 2. Quantitative evaluation of our approach

Criteria	W4 [15]	HoG [2]	Our approach
Precision	0.78	0.95	0.94
Recall	0.70	0.78	0.79
Accuracy	0.76	0.87	0.84
Frame speed	\approx 0.03s	\approx 2.05s	\approx 0.01s

algorithm complexity sits at $O(N)$ where N is the size of each frame, whereas it is $O(N^2)$ in [2]. Note that we developed all the test code using MATLAB environment running on a dual core 64-bit processor @2GHZ with 4GB of RAM. With this setup, we achieved performances of 9 fps (frames of 720x576 pixels).

4 Conclusion and Discussion

In this paper, we presented an efficient and fast method for people detection in videos. The method is based on Adaboost algorithm which operates a binary classification on the silhouettes produced by change detection algorithms. Our experiments demonstrated that our method detects people even in the presence of occlusions and clutter. We demonstrated also that our method is suitable for detecting both isolated individuals or groups of people, which is one of our major contributions. Finally, we showed a quantitative performance for our approach that demonstrate its usefulness and efficiency.

There are several roads leading to our method's improvement. Further work will be done over the background subtraction subject, more specifically on the (now empiric) coefficients C , R , and the thresholding values. Gradient feature extraction from zones other than head/shoulders could be added to improve efficiency. Other improvements could be classification using multi-class Adaboost, to apply this algorithm to other objects (e.g., detecting pedestrians, different types of vehicles, and road obstacles). Finally, we could enrich the algorithm by using other features that aren't related to the silhouette (e.g. motion information).

References

1. Gerónimo, D., López, A., Sappa, A., Graf, T.: Survey of pedestrian detection for advanced driver assistance systems. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 32(7), 1239–1258 (2010)
2. Dalal, N., Triggs, B., Rhone-Alps, I., Montbonnot, F.: Histograms of oriented gradients for human detection. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 886–893 (2005)
3. Viola, P., Jones, M., Snow, D.: Detecting pedestrains using patterns of motion and appearance. In: *IEEE Int'l Conf. on Computer Vision*, pp. 734–741 (2003)
4. Yamauchi, Y., Fujiyochi, H., Iwahori, Y., Kanade, T.: People detection based on co-occurrence of appearance and spatio-temporal features. *Progress in Informatics* 7, 33–42 (2010)

5. Wu, B., Nevatia, R.: Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet part detectors. *Int'l J. of Computer Vision* 75(2), 247–266 (2007)
6. Dollár, P., Wojtek, C., Schiele, B., Perona, P.: People detection: A benchmark. In: *IEEE Conf. on Computer Vision and Pattern Recognition*, pp. 304–311 (2009)
7. Mikolajczyk, K., Schmid, C., Zisserman, A.: Human detection based on a probabilistic assembly of robust part detectors. In: *European Conf. on Computer Vision*, pp. 69–82 (2005)
8. Agarwal, S., Awan, A., Roth, D.: Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 26, 1475–1490 (2004)
9. Papageorgiou, C., Poggio, T.: A trainable system for object detection. *Int'l J. of Computer Vision* 38(1), 15–33 (2000)
10. Allili, M.S., Bouguila, N., Ziou, D.: Finite generalized gaussian mixture modelling and application to image and video foreground segmentation. In: *IEEE Canadian Conf. on Computer and Robot Vision*, pp. 183–190 (2007)
11. Allili, M.S., Bouguila, N., Ziou, D.: Robust video foreground segmentation by using generalized gaussian mixture modeling. In: *IEEE Canadian Conf. on Computer and Robot Vision*, pp. 503–509 (2007)
12. Allili, M.S., Bouguila, N., Ziou, D.: Finite general gaussian mixture modelling and application to image and video foreground segmentation. *J. of Electronic Imaging* 17, 1–13 (2008)
13. Bouwmans, T., El Baf, F., Vachan, B.: Statistical background modeling for foreground detection: A survey. In: *Handbook of Pattern Recognition and Computer Vision* 4 (part 2 chapter 3), pp. 181–199 (2010)
14. Sonka, M., Hlavac, V., Boyle, R.: *Image processing, analysis, and machine vision*. Thompson Learning, Toronto (2008)
15. Haritaoglu, I., Harwood, D., Davis, L.: W4: Real-time surveillance of people and their activities. *IEEE Trans. on Pattern Analysis and Machine Intelligence* 22(8), 809–830 (2000)
16. Freund, Y., Schapire, R.: A decision-theoretic generalization of online learning and an application of boosting. *J. of Computer and System Sciences* 55, 119–139 (1997)
17. Gonzalez, R., Woods, R.: *Digital Image Processing*. Prentice-Hall, Englewood Cliffs (2008)

A Higher-Order Model for Fluid Motion Estimation

Wei Liu and Eraldo Ribeiro

Computer Vision and Bio-Inspired Computing Laboratory,
Florida Institute of Technology, Melbourne, FL 32901, USA
lwei@my.fit.edu, eribeiro@cs.fit.edu,

Abstract. Image-based fluid motion estimation is of interest to science and engineering. Flow-estimation methods often rely on physics-based or spline-based parametric models, as well as on smoothing regularizers. The calculation of physics models can be involved, and commonly used 2nd-order regularizers can be biased towards lower-order flow fields. In this paper, we propose a local parametric model based on a linear combination of complex-domain basis flows, and a resulting global field that is produced by blending together local models using partition-of-unity. We show that the global field can be regularized to an *arbitrary order* without bias towards specific flows. Additionally, the blending approach to fluid-motion estimation is more flexible than competing spline-based methods. We obtained promising results on both synthetic and real fluid data.

Keywords: Fluid-flow estimation, optical flow, holomorphic functions.

1 Introduction

Estimating fluid motion from images is interesting to many science and engineering applications, and has received renewed attention from the computer vision community [2,7]. Fluid-flow estimation differs from the similar optical-flow estimation problem in a number of ways [5]. First, general optical flow fields are often unstructured, while fluid flows usually result from continuous physical processes. As a result, parametric models are common in recent works that produce smooth and accurate results [3]. Secondly, smoothness regularizers in variational optical-flow methods are often based on first-order derivatives [6]. These methods are thus biased towards piecewise-linear flow fields limiting their application to fluid flows. This limitation can be addressed by using a second-order regularizer based on the flow fields' divergence and rotation [2,3,7]. However, it is unclear how higher-order regularizers can be designed. In this paper, we propose a parametric model that is robust to noise, is able to represent complicated turbulence, and has a regularizer that is not biased to lower-order flow fields.

Parametric models of fluid flows can be classified into two main groups. The first group are based on physics priors of fluid dynamics, and integrate temporal information into the motion estimation process, producing temporarily consistent results [5]. However, flow fields described by these models are restricted by physics laws, and, as observed in [7], these methods rely on rather involved minimization processes. The second group of methods do not make explicit use of fluid dynamics, but estimate fluid motion solely based on the apparent image deformation, and rely on simple smoothness heuristics to regularize the estimation results [2,12,7]. This group of methods is

closely related to the classical problem of optical-flow estimation and nonrigid image registration. A recent work by Isambert et al. [7] produced superior results on turbulent flows, using locally supported vector splines, and representing flows using a multi-scale scheme. However, spline-model optimization can be computationally expensive when dense control-point grids are used, and it is sensitive to local minima. On the other hand, the use of sparse control points can oversmooth estimated flow fields. Most importantly, exact minimization of the functional proposed by Isambert et al. [7] leads to thin-plate splines. This means that their model is still biased to certain lower-order flow fields.

To address the above problems, we introduce a simple parametric model, that is robust yet flexible to represent turbulent flow fields, and can be regularized through a convex functional. Our approach belongs to the second group of methods and makes no assumptions about the fluid's physics properties. Similar to [7], we use a locally supported parametric model to represent a flow field. Instead of using splines and interpolating the motion between control points [7], we use a linear model of orthogonal basis functions represented as holomorphic functions, and approximate the global field by blending the local models using partition-of-unity (Section 2). The use of holomorphic models leads to simpler handling of important fluid-flow properties such as divergence and rotation, and allows us to regularize a fluid flow unbiasedly, by penalizing inconsistencies between neighboring local flows instead of their spatial gradients [7,2]. Additionally, the resulting energy functional is convex, and can be minimized through gradient-descent methods (Section 3). We tested our method on motions from both synthetic and real fluid data (Section 4). Finally, we point out the limitations of our holomorphic flow-field model, and directions for future work (Section 5).

2 Higher-Order Model of Flow Field

Parametric models provide a flexible yet compact flow-field representation. In this section, we represent local flow fields using holomorphic complex functions. Local holomorphic models have been previously used to represent singular points in flow fields [8]. Here, we extend these functions to represent both singular and smooth flow regions.

2.1 Local Flow Field Model

We commence by representing a 2-D vector-flow field as a complex-valued function $F(z)$ defined on a finite domain $\Omega \in \mathbb{C}$ [8,11]. This vector-flow field is then approximated by an holomorphic function centered at $z_0 \in \mathbb{C}$, i.e., $f(z) \approx F(z + z_0)$, that can be modeled using a linear combination of complex basis functions (basis flow fields). For example, the Taylor expansion of $f(z)$ about the origin (i.e., $z_0 = 0$) can be written as a linear combination of complex (orthogonal) monomials $\phi_k(z) = z^k$:

$$f(z) = \sum_{k=0}^N a_k \phi_k(z) + R_N(z), \quad (1)$$

where $a_k = \frac{f^{(k)}(0)}{k!}$ are the coefficients, and $R_N(z)$ is the residue. Here, $f^{(k)}(0)$ is the k -th derivative of f evaluated at $z_0 = 0$. For simplicity, we assume the basis $\phi_k(z)$

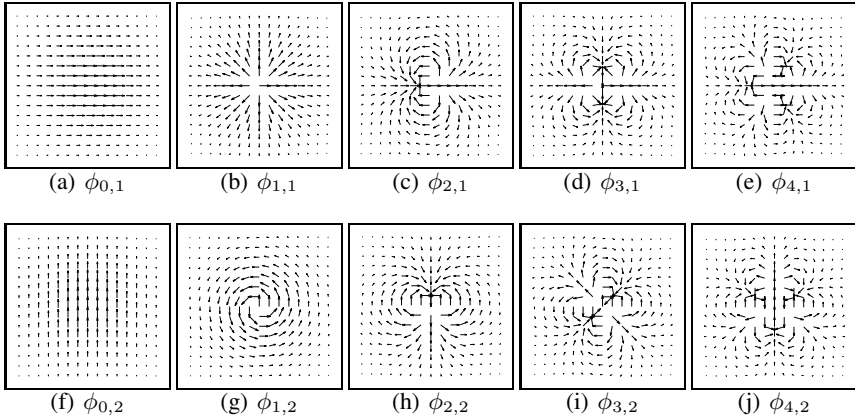


Fig. 1. Basis polynomials $\phi_{k,i}$ multiplied with weight function $w_\sigma(z)$ for $k = 0, \dots, 4$ and $i = 1, 2$. First column: polynomials derived from z^k . Second column: polynomials derived from iz^k . $\phi_{1,1}$ is a rotation-free source field and $\phi_{1,2}$ is a divergence-free vortex. The flow fields exhibit higher-order fluctuation with increasing k .

to be orthogonal, so the coefficients a_k can be calculated by inner product projection. Both the orthogonality condition and projection operator depend on the choice of inner product in the analytic functions space $A(\Omega)$. The classic Hermitian inner product [4] produces complex numbers, making projection calculations difficult. Instead, we use vector fields' correlation [8] as an alternative inner product:

$$\langle f(z), g(z) \rangle = \int_{\mathbb{C}} (f(z) \cdot g(z)) w_\sigma(z) dz, \tag{2}$$

where \cdot is the dot product between two complex numbers, and w_σ is a Gaussian kernel that makes the projection local. Flow-field $f(z)$ can be projected onto the basis function $\phi_k(z)$, with real-domain projection coefficients given by $a_k = \frac{\langle f(z), \phi_k(z) \rangle}{\langle \phi_k(z), \phi_k(z) \rangle}$. Furthermore, we can re-write Equation 2 as:

$$\langle f(z), g(z) \rangle = (F \otimes g)(z_0) = \int_{\mathbb{C}} (F(z + z_0) \cdot g(z)) w_\sigma(z) dz, \tag{3}$$

which can be implemented efficiently using the Fast Fourier Transform (FFT). Given the inner product defined in Equation 2 we can show that complex monomials $\{z^k\}_{k=1}^N$ and $\{iz^k\}_{k=1}^N$ form a complete orthogonal basis. Intuitively, iz^k is a counterclockwise 90-degree rotation of the vectors in z^k . Our basis flows can then be written as: $\phi_{k,1}(z) = z^k$ and $\phi_{k,2}(z) = iz^k$. Figure 1 shows the weighted basis functions $\phi_{k,i} * w_\sigma(z)$ for $k = 0, \dots, 3$. Using (1), the N -th order flow-field approximation at $p \in \Omega$ is:

$$F(z + z_0) \approx f(z) = \sum_{k=0}^N (a_{k,1} \phi_{k,1}(z) + a_{k,2} \phi_{k,2}(z)), \tag{4}$$

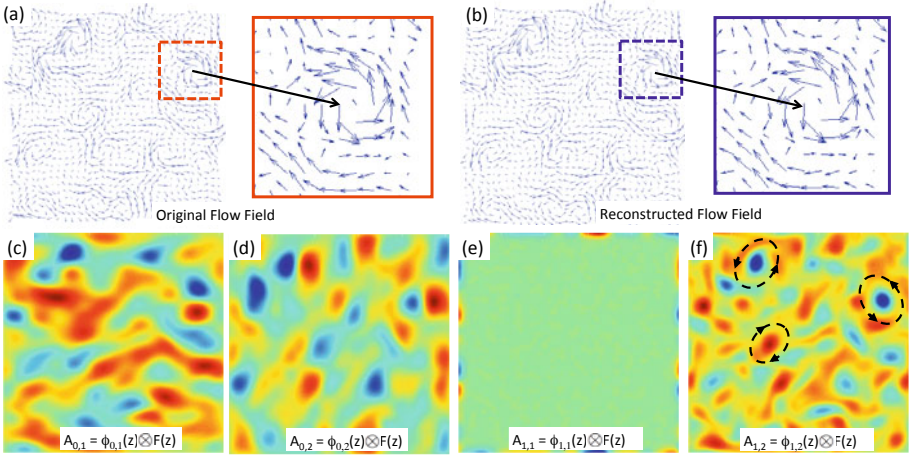


Fig. 2. Decomposition and reconstruction. (a) Original turbulent flow and detail view. (b) Reconstructed flow and detail view. (c)-(f) Correlation coefficient maps for the first four projection coefficients for $\phi_{k,1}(z)$. Coefficient map $A_{1,1}$ in (e) shows that the flow field is divergence free, while stronger responses in $A_{1,2}$ (f) indicate vertex locations. Blue color indicates orientation match between filter and flow data while red indicates reverse orientation.

where $a_{k,i} = \langle f(z), \phi_{k,i}(z) \rangle$, for $k = 1, \dots, N$, and $i = 1, 2$. The approximation produces $2(N + 1)$ real coefficients $a^p = a_{0,1}^p, a_{0,2}^p, \dots, a_{N,1}^p, a_{N,2}^p$ for location p . According to (3), the coefficients are local values of the cross-correlation between $F(z)$ and $\phi_{k,i}(z)$. It can be shown that by letting $z \rightarrow 0$ in Equation 4, the local flow field's divergence and rotation simply equal to $a_{1,1}$ and $a_{1,2}$, respectively. This observation shows that both divergence and rotation are represented in our model. Figure 2 shows the correlation between the first two basis pairs and a turbulent flow, i.e., $A_{k,1} = F(z) \otimes \phi_{k,1}(z)$ and $A_{k,2} = F(z) \otimes \phi_{k,2}(z)$, $k = 0, 1$. The turbulent flow field happens to be divergence free so $A_{1,1}$ vanishes almost everywhere. This further confirms that $a_{1,1}$ and $a_{1,2}$ are related to the divergence and rotation of the flow field.

2.2 Blending Local Models into a Global Flow Field

Local flow models can be blended into a global flow field using a partition-of-unity [7]:

$$\tilde{F}(z) = \sum_{k,i} \int_p A_{k,i}(p) \phi_{k,i}(z-p) h(z-p) dp. \tag{5}$$

Here, function h is a blending function such that $\int h(z) dz = 1$, ensuring that the contributions of neighboring models sum to one (partition-of-unity) [7]. In this paper, we choose $h(z)$ to be a Gaussian function with the same size as our basis flows. This blending approach is more flexible than the interpolating splines [7], as local models are not required to agree at control points. Similarly to splines, the global representation in Equation 5 can be blended using a sparse grid of local models.

3 Fluid Flow Estimation

We now extend the modeling described in previous sections to fluid-flow estimation. In general, fluid-flow estimation is formulated as the following minimization problem [2]:

$$\int D(I(\mathbf{x} + \mathbf{v}, t + \delta t), I(\mathbf{x}, t)) d\mathbf{x} + \lambda \int S(\mathbf{v})d\mathbf{x}, \tag{6}$$

where \mathbf{x} and \mathbf{v} are the spatial and velocity vectors, respectively, D is the data term enforcing luminance or mass constancy, and S is the regularizer preferring smooth solutions. Since luminance constancy simplifies computation, and is widely used for incompressible fluid flows, in this work, we enforce the luminance constancy, and leave the mass constancy for future study. The most common data term used to enforce luminance constancy is based on a quadratic form that can be discretized into the well-known optical-flow constraint as $D(I(\mathbf{x} + \mathbf{v}, t + \delta t), I(\mathbf{x}, t)) = (\nabla I \cdot \mathbf{v} + \partial I/\partial t)^2$, where ∇I is the spatial image gradient, and $\frac{\partial I}{\partial t}$ is the time difference. There are two typical regularizers for regularizing the flow fields, including the first-order Horn-Shunk’s regularizer [6] and the second-order regularizer used in [2], respectively:

$$S^{(1)}(\mathbf{v}) = \|\nabla \mathbf{v}_1\|^2 + \|\nabla \mathbf{v}_2\|^2 \quad \text{and} \quad S^{(2)}(\mathbf{v}) = \|\nabla \text{div}(\mathbf{v})\|^2 + \|\nabla \text{rot}(\mathbf{v})\|^2. \tag{7}$$

$S^{(1)}$ is widely used in optical flow computation, and is biased towards piecewise linear flows, while $S^{(2)}$ is considered more appropriate for regulating fluid motions. Here, since we represent flow fields using parametric models, instead of recovering \mathbf{v} directly, we aim at finding the optimal coefficients representing the underlying motion between two images. In the following section, we first show how the optical-flow constraint and the existing regularizers can be rewritten using the proposed model. Then, we introduce a general regularizer for arbitrary-order flow fields.

3.1 Local Optical-Flow Constraint

Let us write the image gradient as a complex function $\nabla I(z) = \frac{\partial I}{\partial x} + \frac{\partial I}{\partial y}i$, and let $f(z)$ represent \mathbf{v} at pixel p . We can then substitute the linear approximation of $f(z)$ in [4] into the optical-flow constraint to minimize the following weighted error function:

$$D(p) = \sum_{z \in N_p} w_\sigma(z) \left(\sum_{k=0}^N (a_{k,1}\phi_{k,1}(z) + a_{k,2}\phi_{k,2}(z)) \cdot \nabla I(z) - \partial I/\partial t \right)^2, \tag{8}$$

where $w_\sigma(z)$ is a Gaussian function that weights the image evidence more at the center. Equation [8] can be written in a compact matrix form: $(\mathbf{R}_p \mathbf{a}_p - \mathbf{T}_p)^\top \mathbf{W} (\mathbf{R}_p \mathbf{a}_p - \mathbf{T}_p)$, with \mathbf{W} contains the weighting factor $w_\sigma(z)$, \mathbf{R}_p is calculated from $\phi_{k,i} \cdot \nabla I(z)$, and \mathbf{T}_p is obtained by stacking $\frac{\partial I}{\partial t}$. Minimizing Equation [8] leads to a local optical flow calculation similar to the Lucas-Kanade [9] method that can be solved as a linear system.

3.2 Global Smoothness Constraint

In this section, we show how smoothness constraints can be formulated directly from the local coefficients vector $\frac{\partial a^p}{\partial t}, p \in \Omega$. If we consider $z = 0$ in Equation [4], the local

velocity vector is simply $(a_{0,1}, a_{0,2})$. Thus, the first-order regularizer can be written as $S^{(1)}(\mathbf{v}) = \|\nabla a_{0,1}\|^2 + \|\nabla a_{0,2}\|^2$. Furthermore, we have shown in Section 2.1 that $\mathbf{div} f(z) = a_{1,1}$ and $\mathbf{rot} f(z) = a_{1,2}$ when $z \rightarrow 0$. As a result, the second-order regularizer becomes: $S^{(2)}(\mathbf{v}) = \|\nabla a_{1,1}\|^2 + \|\nabla a_{1,2}\|^2$. Similarly, we can define an arbitrary-order regularizer $S_1^N = \sum_{k=0}^N \beta_k \left(\|\nabla a_{k,1}\|^2 + \|\nabla a_{k,2}\|^2 \right)$ where $\beta_k \geq 0$ are weight factors that emphasize on different orders, or equivalently:

$$S_1^{(n)} = \sum_{z \in N_p} (\mathbf{a}_z - \mathbf{a}_p)^\top \mathbf{\Gamma} (\mathbf{a}_z - \mathbf{a}_p), \quad (9)$$

where $\mathbf{\Gamma} = \text{diag}(\beta_0, \dots, \beta_N)$, and N_p is the set of neighboring local models. By choosing small β_k for lower-order coefficients, we avoid penalizing lower-order variations. This can be justified by noticing that $a_{k,i}$ is related to the flow field's derivatives through Taylor's expansion in (1), and the n -th order spatial derivatives of $F(z)$ can be measured from the derivatives of the corresponding coefficients $a_{n,i}, i = 1, 2$.

Unbiased higher-order regularizer. The regularizer in (9) penalizes spatial variations of model parameters, and is similar to the one used in (10). However, penalizing model parameters' gradients may lead to bias towards certain orders of flow fields, depending on the choice of the weighting parameters $\mathbf{\Gamma}$. Also, simply penalizing model parameters' spatial variations ignores the fact that the variation can be partially caused by local coordinate system shifting. For example, the local flow $f(z) = z + z^2$ observed at a neighboring position $z + \delta z$ will be $f(z + \delta z) = (\delta z)^2 + (1 + 2\delta z)z + z^2$. In other words, there will be model parameter variations even when the flow field follows exactly a polynomial model. We account for these variations by shifting the local parameters before comparison with neighboring models. Fortunately, we can write the shifting of a basis function (monomial) z^k as a linear combination of lower-order monomials, i.e., $(z + \delta z)^k = (\delta z)^k + (\delta z)^{k-1}z + \dots + z^k$. As result, the shifting operator can be written as a lower-triangular matrix $H(\delta z)$. Thus, an alternative regularizer can be defined as:

$$S_2^{(n)} = \sum_{z \in N_p} (\mathbf{H}(p - z)\mathbf{a}_z - \mathbf{a}_p)^\top \mathbf{\Gamma} (\mathbf{H}(p - z)\mathbf{a}_z - \mathbf{a}_p), \quad (10)$$

where $\mathbf{H}(p - z)$ is the shifting matrix, and the weighting matrix $\mathbf{\Gamma}$ is used to make the notation consistent with Equation 9. In this paper, we simply choose $\mathbf{\Gamma} = \lambda \mathbf{I}$ with $\lambda > 0$, for both (9) and (10). In this way, we are not penalizing the spatial variations of the model parameters. Instead, we penalize the *inconsistency* between local models, so flow fields with different orders will not be biased by the regularizer for the *magnitude* of their variations, as long as they make *consistent* variations. It is easy to verify that any holomorphic functions (flows) with order less than N can make $S_2^{(n)}$ vanish, and this confirms that lower- and higher-order flow fields are equally penalized.

3.3 Gradient-Descent Minimization

We now combine both local and global constraints into a single functional as:

$$E_1^{(n)} = \sum_{p \in \Omega} D(p) + S_1^{(n)}(p) \quad \text{or} \quad E_2^{(n)} = \sum_{p \in \Omega} D(p) + S_2^{(n)}(p). \quad (11)$$

Here, both $E_1^{(n)}$ and $E_2^{(n)}$ are convex, and can be minimized using variational calculus. Since their minimizing procedures are analogous, we will only explain the minimization for $E_2^{(n)}$. The gradients for this functional can be derived as follows:

$$\frac{\partial E^N}{\partial \mathbf{a}_p} = 2 \left\{ \mathbf{a}_p^\top \underbrace{(\mathbf{R}_p^\top \mathbf{W} \mathbf{R}_p + \|N_p\| \Gamma)}_{\mathbf{M}_p} - \sum_{z \in N_p} \mathbf{a}_z^\top \underbrace{\mathbf{H}^\top (p - z) \Gamma}_{\text{shifting term}} - \underbrace{\mathbf{T}_p^\top \mathbf{W} \mathbf{R}_p}_{\mathbf{N}_p} \right\}. \quad (12)$$

In (12), matrices \mathbf{M}_p and \mathbf{N}_p is pre-calculated from image gradients and basis flows. The same applies to the shifting term $\mathbf{H}^\top (p - z) \Gamma$.

4 Experiments

The goal of our experiments is to show that fluid-motion estimation can be improved using our high-order model. We began by evaluating the homomorphic model by obtaining decompositions and reconstructions on synthetic turbulent flows. Then, we ran our fluid-motion estimation algorithms on both synthetic and real images. In all implementations, we used luminance-constancy instead of mass-constancy constraint. The reconstruction's average end-point error (APE) on European FLUID dataset [1] using 2nd-order and 3rd-order models as a function of basis-flow radius was less than 5%, showing that the fluid motion was well represented by our model. It is worth noticing that as the radius of the local models approaches zero, our representation becomes over-parameterized, and the 3rd-order model produced larger reconstruction error for the radius were smaller than two pixels.

Synthetic PIV images. On synthetic images, we quantitatively compared the following methods: the classic Horn-Shunk method [1] ($S^{(1)}$), a B-spline adaptation of the method in [7], and also with our higher-order regularizer without shifting ($E_1^{(n)}$) and with shifting ($E_2^{(n)}$). For the last two, we tested the cases of $n = 2$ and $n = 3$. Although we do not have implementations of the second-order regularizer used in [2] (i.e., $S^{(2)}$), our regularizer $E_1^{(2)}$ can be seen as a parametric version of it. As ground truth is hard to obtain for fluid images, we resorted to synthetic PIV images from the FET-Open European project FLUID [1]. This database contains 6 different types of stable flows, and turbulent flows. As stable and turbulent flows are different in nature, we tuned the algorithm parameters separately for each dataset. These parameters are: (1) the smoothness weight λ_{hs} for Horn-Shunk's method; (2) the spacing of control points d_{sp} , and the smoothness weight λ_{sp} for spline-based method; (3) for our method, the scale (radius) of the parameterized model r , the spacing between local models d , and the regularizer weight λ . Table 1 summarizes the parameters used for each method and dataset.

Tables 2 show the average angular error (AAE) and average end-point error (APE) of the compared methods. Our method performed better on almost all sequences, and the errors decrease with increasing approximation order. Comparing results of $E_1^{(n)}$ and $E_2^{(n)}$ shows that the shifting operator increases estimation accuracy. Additionally, the

¹ Available for download from: <http://www.cs.brown.edu/~dqsun/>

Table 1. Algorithm Parameters

Dataset	Horn-Shunk	Spline		$E_1^{(2)} & E_2^{(2)}$			$E_1^{(3)} & E_2^{(3)}$		
	λ_{hs}	λ_{sp}	d_{sp}	λ	d	r	λ	d	r
Stable Flows	2500	0.1	32	0.1	8	32	0.5	8	32
Turbulence	1500	0.1	8	0.1	2	6	0.5	2	6

Table 2. AAE and APE on Analytic Fluid Sequence

	Seq. 1		Seq. 2		Seq. 3		Seq. 4		Seq. 5		Seq. 6		Turb.	
	AAE	APE	AAE	APE	AAE	APE	AAE	APE	AAE	APE	AAE	APE	AAE	APE
HS	1.02	0.04	1.96	0.04	1.01	0.04	2.75	0.06	2.77	0.06	1.62	0.05	22.09	0.43
Spline	0.63	0.03	0.96	0.02	1.13	0.04	2.73	0.06	2.28	0.05	1.43	0.05	7.27	0.13
$E_1^{(2)}$	0.85	0.04	1.70	0.04	0.78	0.03	2.48	0.05	2.59	0.05	1.34	0.04	4.62	0.09
$E_2^{(2)}$	0.80	0.03	1.63	0.03	0.72	0.03	2.43	0.05	2.53	0.05	1.31	0.04	4.62	0.08
$E_1^{(3)}$	0.58	0.03	1.38	0.03	0.63	0.03	1.87	0.04	1.88	0.04	1.45	0.06	4.58	0.08
$E_2^{(3)}$	0.58	0.03	1.24	0.02	0.55	0.02	1.87	0.04	1.89	0.04	1.49	0.18	4.30	0.08

experiments confirmed the observation in [7] that spline-based methods produce better results than their nonparametric counterparts, especially on turbulent flows. Figure 4 shows streamline and vorticity maps of the extracted turbulence motion. Although the difference is visually small from the streamlines, it can be seen that both the spline-based and our method produce a ‘smoother’ vorticity map than the Horn-Shunk method, and that our method’s vorticity map is closest to the ground truth in its magnitude.

Real-world images. In Figure 4 we show the estimated motion from a wingtip vortex [3] and satellite images [3]. Both the spline model and ours produce smoother results than the Horn-Shunk method. However the spline model easily got trapped in local minima when small smoothness parameters were used, and produced over-smoothed results when the parameter was large. Specifically, for satellite images, all the three methods produced a weak flow for static image regions due to the smoothness constraint. Interestingly, as we have discussed in Section 3.2 Horn-Shunk’s first-order regularizer produced piecewise linear flows, and the spline model split the flow field to satisfy the thin-plate deformation energy, while ours produced consistent background flow.

5 Limitations of Our Method and Future Work

We have proposed a higher-order model of flow fields using complex polynomials. Using this model, we were able to reformulate the optical flow computation in a general

² Courtesy of ONERA

³ Copyright @ EUMETSAT

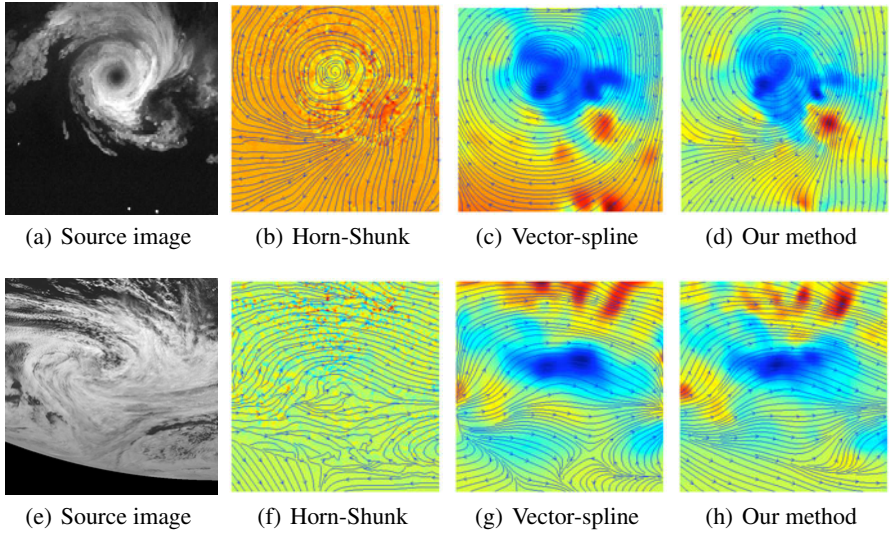


Fig. 3. Real-world image sequences. The first row shows flow fields estimated from a Wingtip Vortex, and the second row shows the ones from satellite images. Compared to the spline model, our method does not over-smooth the flow fields, and produce more consistent results.

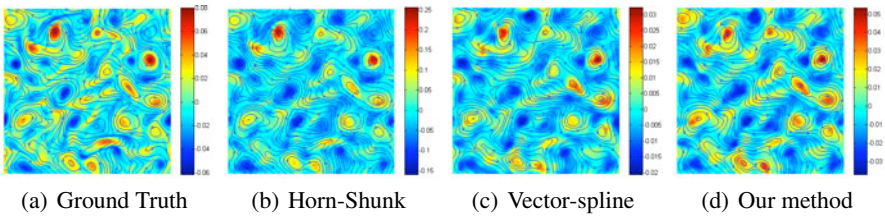


Fig. 4. Fluid motion estimation. Both spline-based methods and ours produced smoother results than Horn-Shunk’s. The vorticity estimated by our method is closer to the ground truth.

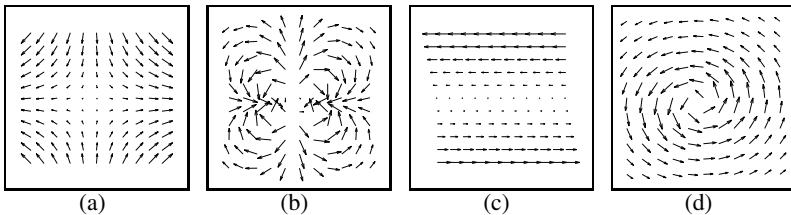


Fig. 5. Flow fields that cannot be well approximated by holomorphic functions of similar scales, including a conjugate flow $f(z) = \bar{z}$ (a) and its holomorphic approximation shown in (b), a shear flow $f(z) = z + \bar{z}$ (c) and its holomorphic approximation shown in (d)

way in which the regularizer can be chosen to penalize certain orders of variations. It is important to point out that the holomorphic assumption used in our approximation model is restrictive as certain flow fields may not be well represented by our model.

Figure 5 shows two examples of such flows, namely, the conjugate flow, $f(z) = \bar{z}$, and the affine flow, $f(z) = z + \bar{z}$, with their holomorphic approximations using basis flows of similar scales to the approximated local flows. Both of the flows are non-analytic anywhere in the complex plane, and their holomorphic approximations are poor. This problem can be partially addressed by minimizing the basis flows' scales. In the extreme case when the bases' scale approaches zero, our flow-field model become over-parameterized, and the flow fields can be fully represented. However, this would increase computational cost, and we believe the better solution lies in extending our approximation model to include non-analytic basis flows. Our future work also includes extension of the method to 3-D flow-field estimation, integration with flow-field singular pattern detection [8], and the usage of mass-constancy constraints [2].

References

1. Carlier, J.: Second set of fluid mechanics image sequences. European Project 'Fluid image analysis and description', FLUID (2005), <http://www.fluid.irisa.fr/>
2. Corpetti, T., Memin, E., Prez, P.: Dense estimation of fluid flows. *IEEE Trans. Patt. Anal. and Mach. Intel.* 24(3), 365–380 (2002)
3. Cuzol, A., Hellier, P., Memin, E.: A low dimensional fluid motion estimator. *International Journal of Computer Vision* 75(3), 329–349 (2007)
4. Davies, B.: *Integral Transforms and Their Applications*. Springer, Heidelberg (2002)
5. Heitz, D., Memin, E., Schnorr, C.: Variational fluid flow measurements from image sequences: synopsis and perspectives. *Experiments in Fluids* 48(3), 369–393 (2009)
6. Horn, B., Schunck, B.: Determining optical flow. *Artificial Intelligence* 17(1-3), 185–203 (1981)
7. Isambert, T., Berroir, J.-P., Herlin, I.: A multi-scale vector spline method for estimating the fluids motion on satellite images. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part IV. LNCS*, vol. 5305, pp. 665–676. Springer, Heidelberg (2008)
8. Liu, W., Ribeiro, E.: Scale and Rotation Invariant Detection of Singular Patterns in Vector Flow Fields. In: *S-SSPR*, pp. 522–531 (2010)
9. Lucas, B.D., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: *IJCAI*, pp. 674–679 (1981)
10. Nir, T., Bruckstein, A.M., Kimmel, R.: Over-parameterized variational optical flow. *Int. J. Comput. Vision* 76(2), 205–216 (2008)
11. Petrla, T., Trif, D.: *Basics of fluid mechanics and introduction to computational fluid dynamics*. Springer, Heidelberg (2005)
12. Suter, D.: Motion estimation and vector splines. In: *CVPR*, pp. 939–942 (2002)

Dictionary Learning in Texture Classification

Mehrdad J. Gangeh¹, Ali Ghodsi², and Mohamed S. Kamel¹

¹ Pattern Analysis and Machine Intelligence (PAMI) Lab,
Department of Electrical and Computer Engineering, University of Waterloo,
200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1
{mgangeh, mkamel}@pami.uwaterloo.ca

² Department of Statistics and Actuarial Science, University of Waterloo,
200 University Avenue West, Waterloo, Ontario, Canada N2L 3G1
agheidsib@uwaterloo.ca

Abstract. Texture analysis is used in numerous applications in various fields. There have been many different approaches/techniques in the literature for texture analysis among which the texton-based approach that computes the primitive elements representing textures using k -means algorithm has shown great success. Recently, dictionary learning and sparse coding has provided state-of-the-art results in various applications. With recent advances in computing the dictionary and sparse coefficients using fast algorithms, it is possible to use these techniques to learn the primitive elements and histogram of them to represent textures. In this paper, online learning is used as fast implementation of sparse coding for texture classification. The results show similar to or better performance than texton based approach on CURET database despite of computation of dictionary without taking into account the class labels.

Keywords: Dictionary learning, matrix factorization, sparse coding, texture classification.

1 Introduction

Texture provides important information in various fields of image analysis and computer vision. It has been used in many different problems including texture classification, texture segmentation, texture synthesis, material recognition, 3D shape reconstruction, color-texture analysis, appearance modeling, and indexing [1-4].

As texture is a complicated phenomenon, there is no definition that is agreed upon by the researchers in the field [2, 3]. This is one of the reasons that there are various analysis techniques in the literature, each of which tries to model one or several properties of texture depending on the application in hand.

Among these techniques, the approaches based on representing textures using some primitive elements, either predefined or learned, has recently shown great success in texture analysis. These approaches have roots in influential paper by Julesz [5]. He introduced textons as fundamental primitive elements that can describe

textures. However, he did not propose any method how to compute these primitive elements in [5].

Based on Julesz proposal, two techniques have recently obtained prevalence in texture analysis. First, techniques based on local binary patterns (LBPs) [6], in which *fixed* operators, i.e., LBPs and their histogram are used to represent a texture. Second, texton-based approach where *learned* textons (composed in a dictionary) are used as primitive elements to represent a texture. Our focus in this paper is on this latter approach, i.e., *learned* dictionary of textons.

Leung and Malik were the first to develop a complete texture classification system using texton-based approach [7]. They defined 2D textons as the cluster centers in filter bank responses, which made it possible to generate textons from the images automatically as the prototypes representing the source textures. These textons formed a dictionary from which a texton histogram could be constructed for each image using a similarity measure. Their work was further improved by Schmid [8], Cula and Dana [9], and Varma and Zisserman [10, 11].

In texton-based approach, the textons in the dictionary are learned using a clustering algorithm such as k -means. However, as explained in [10], one main shortcoming of k -means is that it can be only applied to points within a texture class. It cannot be applied across classes as it merges data points (by taking mean of points) and thus the resultant cluster centers cannot be identified uniquely with individual textures. This means that the cluster centers computed using k -means across classes are not representing textures in a class anymore.

A solution to this problem is computing the dictionary using dictionary learning approaches based on sparse coding or using matrix factorization¹. Previously, these approaches for dictionary learning were too slow to be utilized in these applications. However, with recent advances in this field and by introducing fast algorithms such as online learning [12], rank-one downdate (R1D) [13], and coordinate descent [14], it is now computationally feasible to compute the dictionary on millions of patches (data samples in general) in reasonable time. This means that the dictionary can be learned on whole training set (not per class) using these approaches. The main advantage is that we do not use the class labels at this stage, i.e., learning dictionary is fully unsupervised.

Here, we propose using online learning [12] for learning a dictionary on the whole training set and computation of sparse coefficients over the whole dictionary and show that despite of fully unsupervised learning of dictionary, on standard databases such as Columbia Utrecht Reflectance and Texture (CURET) database [15], it performs similar to or better than texton-based approaches using k -means, where dictionary is learned per class.

The rest of the paper is organized as follows: Section 2 presents the theory of dictionary learning and sparse coding (DLSC) related to our work. Experimental setup is described in Section 3 followed by results in Section 4. The paper is concluded in Section 5.

¹ The connection between matrix factorization and dictionary learning using sparse coding is explained in [12].

2 Dictionary Learning and Sparse Coding

In this section, we first provide an overview of dictionary learning and sparse coding (DLSC) and its connection to texton-based approach for texture classification. Then we provide the formulation for dictionary learning with sparse representation for texture classification.

2.1 Background

Dictionary learning and sparse representation/coding are two closely related topics in the literature. The initial work on these two topics was originated from two communities and problems under two different names, i.e., sparse coding (SC), which was originated by neurologists as a model for simple cells in mammalian primary visual cortex [16, 17]; and, independent component analysis (ICA), which was originated by researchers in signal processing to estimate the underlying hidden components of multivariate statistical data (refer to [18] for a review of ICA). These two problems merged, eventually, into similar techniques, but somewhat different description (the connection between SC and ICA is also explained in [18]).

The main result of these two research works was that a class of signals with sparse nature, such as the images of natural scenes, can be represented using some primitive elements that form a dictionary, and that each signal in this class, can be represented by using only few elements in the dictionary (sparse representation).

In fact, there are, at least, two ways in the literature to exploit sparsity [19]: first, using a linear/nonlinear combination of some *predefined* bases, e.g., wavelets [20]. Second, by using primitive elements in a *learned* dictionary, such as techniques employed in SC or ICA. This latter approach is our focus in this paper.

As mentioned in the introduction, dictionary learning was introduced to the field of texture analysis by Julesz theory that stated textures can be represented using a few primitive elements [5] and following the work done in [7, 8, 9, 10, 11] that initiated the texton-based approach in texture classification. Texton-based approach mainly consists of two steps, dictionary learning and computation of models (features) for each texture image. In the first step, extracted patches from each texture image in a class are submitted to a clustering algorithm such as *k*-means and obtained cluster centers are used as primitive elements (called textons) that form the dictionary. In the second step, for each texture image, a histogram of textons is computed. To compute this histogram, patches are extracted from each texture image and each patch is compared with the textons in the dictionary. The closest match based on a similarity measure such as Euclidean distance is used to update the corresponding bin in the histogram of textons. Thus, each patch in a texture image is represented by only one single texton in the dictionary (the closest match). This is a kind of sparse representation, in which only one atom in the dictionary is active per patch. These two steps can be performed using DLSC, which is described next.

2.2 Mathematical Formulation

Considering a finite training set of signals $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m] \in \mathbb{R}^{d \times m}$, they can be represented by a dictionary \mathbf{D} and a set of sparse coefficients α using

$$\min_{\mathbf{D}, \alpha} \sum_{i=1}^m \left(\frac{1}{2} \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda \varphi(\alpha_i) \right), \quad (1)$$

where λ is a regularization parameter and $\varphi(\cdot)$ is a sparsity inducing function. The most common sparsity inducing function is ℓ_1 norm and the corresponding problem is known as the *Lasso* [21]

$$\min_{\mathbf{D}, \alpha} \sum_{i=1}^m \left(\frac{1}{2} \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1 \right). \quad (2)$$

To prevent obtaining very large values of \mathbf{D} , which consequently leads to very small values of α_i , a constraint is imposed on the columns of \mathbf{D} such that they have unit ℓ_2 norm [12].

Solving (2) using one of the approaches in the literature such as online learning [12] yields the dictionary \mathbf{D} and the sparse coefficients α . If the dictionary has been already computed (using all $\mathbf{x}_i, i = 1, \dots, m$ in the training set), (2) can be used to find the sparse coefficients for a signal \mathbf{x} in test set (\mathbf{D} is fixed in this case).

2.3 Texture Classification

Texture classification using dictionary learning and sparse representation based on (2) can be done in two steps. In first step, the dictionary $\mathbf{D} \in \mathbb{R}^{d \times k}$ (k is the number of primitive elements in the dictionary) is learned using $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_m] \in \mathbb{R}^{d \times m}$, where $\mathbf{x}_i, i = 1, \dots, m$ are patches extracted with size $\sqrt{d} \times \sqrt{d}$ from texture images in training set in all classes. With fast algorithms such as R1D or online learning, this can be performed in few minutes over millions of patches.

After learning the dictionary, we need to find the model (feature set) for each texture image in training and test sets. To this end, patches of the same size as what is used in dictionary learning step are extracted from each texture image, i.e., $\mathbf{X} = [\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n] \in \mathbb{R}^{d \times n}$, where n is the number of patches extracted, which is not necessarily the same as m . Then using (2), the corresponding coefficients $\alpha_i \in \mathbb{R}^{k \times n}, i = 1, \dots, n$ are computed. For each patch \mathbf{x}_i , most of the elements in the corresponding coefficient α_i are zero. The nonzero elements in α_i determine the primitive elements in the dictionary \mathbf{D} that contribute towards the representation of the patch \mathbf{x}_i . If we sum up all these coefficients for all patches extracted from a texture image, we effectively find the histogram of primitive elements contributing towards the representation of this particular texture, i.e.,

$$\mathbf{H}(\mathbf{X}) = \sum_{i=1}^n \alpha_i. \quad (3)$$

We impose a positive constraint on α_i in (2) such that we eventually obtain a histogram \mathbf{H} with positive values in all bins. This also prevents cancelling the effect of different patches when they are summed up in (3). Hence, we rewrite (2) as follows to consider this constraint as well as the constraint we considered on \mathbf{D} columns in previous subsection

$$\begin{aligned} \min_{\mathbf{D}, \alpha} \sum_{i=1}^m \left(\frac{1}{2} \|\mathbf{x}_i - \mathbf{D}\alpha_i\|_2^2 + \lambda \|\alpha_i\|_1 \right), \\ \text{s.t. } \forall j = 1, \dots, k, \quad \|\mathbf{d}_j\|_2 = 1 \quad \& \quad \alpha_i \geq 0, \end{aligned} \quad (4)$$

where \mathbf{d}_j is the j^{th} column of \mathbf{D} . In this way, while in texton-based approach each patch is represented using only the closest texton in the dictionary, here each patch is represented by using several primitive elements in the dictionary and hence it can potentially provide richer representation than texton-based approach. The number of nonzero elements in α_i can be controlled using λ in (4), i.e., larger values of λ yield sparser coefficients [12].

The distance between two normalized histograms is measured using χ^2 statistic, i.e., using $\chi^2(H_1, H_2) = 1/2 \sum_i (h_{1i} - h_{2i})^2 / (h_{1i} + h_{2i})$. One nearest neighbor is used as the classifier as suggested in [11].

Although, dictionary learning is also used in [22] for texture classification, our work is different in following three aspects. Firstly, in [22] one dictionary is learned per class and then these dictionaries are composed (concatenated) to form the overall dictionary (this is the same as what is reported in the literature for finding dictionary using k -means). We find the dictionary on the whole training set (not per class) and this means that we do not use class labels at this stage at all. Secondly, to find the sparse coefficients, in [22] part of dictionary which is most similar to the current patch is considered (it is not explained what kind of similarity is used) and the reason mentioned is that using the whole dictionary is computationally very expensive. We find the sparse coefficients on whole dictionary (this is possible with recent advances in computation of the *Lasso* in the literature as mentioned before). Thirdly, we have placed positive constraint on the coefficients as we eventually sum them up to find the histogram of primitive elements (in the dictionary) as the feature set for an image to be classified. In [22] this positive constraint on the coefficients is not considered and this might not be needed as the coefficients are not found on the whole dictionary but just on part of dictionary most similar to the current patch. In fact, our experiments show that without this positive constraint on the coefficients, the performance of the classification system is very poor.

3 Experimental Setup

The performance of the proposed classification system is evaluated on CURET database. The database is used the same as what is reported in [11]. That is, there are 92 images per class and 61 classes. Each image is 200×200 pixels with the intensity resolution of 8 bit/pixel. The comparison is made with texton-based approach using raw pixel representation. This means that no filter banks are used.

Data Preparation and Preprocessing. To make the images indiscriminable to the average intensity level and contrast, the mean of texture images is removed and they are also normalized to have unit standard deviation.

Computation of Dictionary. To compute the dictionary, 500 random patches are extracted from each texture image in the training set. Patch sizes of 5×5 , 7×7 , and 9×9 are used in the experiments. No filter banks are applied and raw pixel representation is used. The mean of patches are removed to make the images locally invariant to the average intensity. In texon-based approach, Weber's law normalization is used as reported in [10, 11]. In DLSC, each patch is normalized to have unit ℓ_2 norm. This is done based on the constraint on primitive elements in the dictionary as stated in (4). In texon-based approach, all patches belonging to one class are submitted to the k -means algorithm to find the cluster centers. These cluster centers over all classes are then composed into a single dictionary. In DLSC approach, all patches from all classes are used at once for learning the dictionary. Hence, no class labels are used at this stage. Online learning [12] is used for the implementation of (4) in DLSC. As suggested in [12], the regularization parameter λ in (4) is chosen as $1.2/\sqrt{d}$, where $d = (\text{patch size})^2$. This yields about 10 nonzero coefficients in average for the patches of 9×9 .

Learning Models (Histograms). After computation of the dictionary, we need to find the model. To this end, small overlapping patches with the same size as what was used in the previous step are extracted from the top left to the bottom right of each ROI. As in the dictionary learning, no filter bank is used and raw pixel representation is considered. The mean of each patch is removed and they are normalized according to Weber's Law in texon-based approach and to unit ℓ_2 norm in DLSC. In texon-based approach, Euclidean distance is used as the similarity measure to find the closest texon in the dictionary to each patch. In DLSC, online learning implementation of (4) is used with the same λ value as previous step with positive constraint on the coefficients α . Each coefficient is normalized to sum to one and all of them are then summed up to yield the overall frequency histogram of primitive elements for each texture image, which is used as the signature (model) of the particular texture image after normalization.

4 Results

In this section, we present the results of texture classification on CURET database using both texon-based and DLSC approaches.

Fig. 1 compares the dictionary learned using these two techniques. In texon-based approach, 10 textons are learned in each class using k -means and eventually all textons are composed into a dictionary (610 textons for 61 classes). As can be seen in Fig. 1, every 10 adjacent textons are similar as they are taken from the same class. In DLSC approach, all patches extracted from all classes are used for the learning of dictionary using (4). Hence class labels are not used at this stage. Different from texon-based dictionary, the primitive elements from all classes are spread over entire dictionary in DLSC.

Table 1 shows the performance of one nearest neighbor classifier using texon-based and DLSC approaches. The experiments are repeated 100 times over random sets of training and test sets. The performance is compared for three different patch and four different training set sizes. As can be seen from this table, the performance

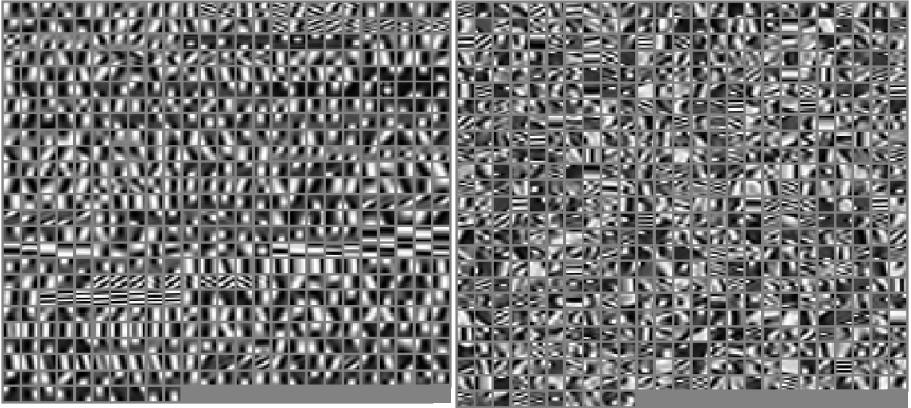


Fig. 1. Dictionary of 610 primitive elements learned using patches of size 7×7 extracted from 23 training texture images per class using: (left) *k*-means algorithm where 10 textons per class are learned and all these textons are composed into a dictionary and (right) DLSC as described in this paper where all primitive elements are learned at once by submitting all extracted patches from all classes to (4).

Table 1. Comparison between the classification accuracy of texton-based and DLSC approaches. The experiments are repeated 100 times on various random split of training and test sets. The dictionary is consisting of 610 primitive elements and results are reported for different train and patch sizes.

Patch Size \ Train Size	5×5		7×7		9×9	
	Texton	DLSC	Texton	DLSC	Texton	DLSC
6	73.76 ± 4.25	74.22 ± 4.37	74.73 ± 4.15	75.77 ± 4.27	75.65 ± 3.92	76.32 ± 4.14
12	83.46 ± 2.60	84.02 ± 2.55	84.25 ± 2.66	85.03 ± 2.55	85.20 ± 2.54	85.32 ± 2.49
23	90.09 ± 1.56	90.52 ± 1.55	90.81 ± 1.62	91.33 ± 1.57	91.42 ± 1.61	91.62 ± 1.59
46	94.83 ± 0.95	95.26 ± 0.93	95.49 ± 0.93	95.85 ± 0.87	95.94 ± 0.85	96.14 ± 0.87

of DLSC is similar to or better than texton-based approach in all cases. This is while the dictionary of DLSC is learned over whole training set at once whereas dictionary of texton-based approach is learned per class, i.e., class labels are taken into account in this learning.

5 Discussion and Conclusion

Sparse representation using few primitive elements learned from data has recently shown great success in different fields such as face recognition and denoising. One of main obstacles for widespread application of this approach was rather slow algorithms

for the computation of dictionary and sparse coefficients over millions of data samples, which is usually the case in image processing and computer vision tasks. The initial algorithm proposed in [16], for example, took hours to compute the dictionary over patches extracted from only ten natural scenes.

With recent fast algorithms proposed for dictionary learning and sparse coding such as online learning and RID, it is now feasible to perform the computation over millions patches in few minutes. In this paper, we proposed using one of these algorithms, i.e., online learning, for the purpose of texture classification over large databases such as CURET. In contrast to k -means algorithm used in texon-based approach that has to learn the dictionary per class, the proposed approach can learn the dictionary over all classes and hence class labels are not used at all in this step. Yet, the results of classification are similar to or better than texon-based approach. The positive constraint imposed on sparse coefficients enables learning the coefficients over whole dictionary and, consequently, finding the model histogram for each texture image is as simple as summing up the sparse coefficients learned for all patches extracted from the particular texture image.

In future work, we would also like to impose positive constraint on the dictionary and utilize nonnegative matrix factorization using fast implementations such as RID [13]. We would also like to extend this work to supervised dictionary learning [19] and compare it to our current results for possible further improvements.

Acknowledgments. The first author gratefully acknowledges the funding from the Natural Sciences and Engineering Research Council (NSERC) of Canada under Canada Graduate Scholarship (CGS D3-378361-2009).

References

1. Petrou, M., Sevilla, P.G.: *Image Processing Dealing with Texture*. John Wiley and Sons, West Sussex (2006)
2. Ahonen, T., Pietikainen, M.: Image Description Using Joint Distribution of Filter Bank Responses. *Pattern Recognition Letters* 30(4), 368–376 (2009)
3. Mirmehdi, M., Xie, X., Suri, J.: *Handbook of Texture Analysis*. Imperial Collage Press, London (2008)
4. Hadjidemetriou, E., Grossberg, M.D., Nayar, S.K.: Multiresolution Histograms and Their Use for Recognition. *IEEE Trans. on PAMI* 26(7), 831–847 (2004)
5. Julesz, B.: Textons, the Elements of Texture Perception, and Their Interactions. *Nature* 290(5802), 91–97 (1981)
6. Ojala, T., Pietikainen, M., Maenpaa, T.: Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. on PAMI* 24(7), 971–987 (2002)
7. Leung, T., Malik, J.: Representing and Recognizing the Visual Appearance of Materials Using Three-Dimensional Textons. *Int'l J. Computer Vision* 43(1), 29–44 (2001)
8. Schmid, C.: Weakly Supervised Learning of Visual Models and Its Application to Content-Based Retrieval. *International Journal of Computer Vision* 56(1/2), 7–16 (2004)
9. Cula, O.G., Dana, K.J.: 3D Texture Recognition Using Bidirectional Feature Histograms. *International Journal of Computer Vision* 59(1), 33–60 (2004)

10. Varma, M., Zisserman, A.: A Statistical Approach to Texture Classification from Single Images. *International Journal of Computer Vision: Special Issue on Texture Analysis and Synthesis* 62(1-2), 61–81 (2005)
11. Varma, M., Zisserman, A.: A Statistical Approach to Material Classification Using Image Patch Exemplars. *IEEE Trans. on PAMI* 31(11), 2032–2047 (2009)
12. Marial, J., Bach, F., Ponce, J., Sapiro, G.: Online Learning for Matrix Factorization and Sparse Coding. *Journal of Machine Learning Research* 11, 19–60 (2010)
13. Biggs, M., Ghodsi, A., Vavasis, S.: Nonnegative Matrix Factorization via Rank-One Downdate. In: *Int'l Conf. on Machine Learning (ICML)*, Helsinki, Finland, pp. 64–71 (2008)
14. Friedman, J., Hastie, T., Tibshirani, R.: Regularized Paths for Generalized Linear Models via Coordinate Descent. *Journal of Statistical Software* 33(1), 1–22 (2010)
15. Dana, K.J., van Ginneken, B., Nayar, S.K., Koenderink, J.J.: Reflectance and Texture of Real-World Surfaces. *ACM Transactions on Graphics* 18(1), 1–34 (1999)
16. Olshausen, B.A., Field, D.J.: Emergence of Simple-Cell Receptive Field Properties by Learning a Sparse Code for Natural Images. *Nature* 381, 607–609 (1996)
17. Olshausen, B.A., Field, D.J.: Sparse Coding with an Overcomplete Basis Set: A Strategy Employed by V1? *Vision Research* 37(23), 3311–3325 (1997)
18. Hyvärinen, A., Karhunen, J., Oja, E.: *Independent Component Analysis*. John Wiley and Sons, New York (2001)
19. Marial, J., Bach, F., Ponce, J., Sapiro, G., Zisserman, A.: Supervised Dictionary Learning. In: *22nd Conference on Neural Information Processing Systems (NIPS)*, Vancouver, Canada, pp. 1033–1040 (2008)
20. Mallat, S.: *Wavelet Tour of Signal Processing: The Sparse Way*, 3rd edn. Academic Press, Burlington (2009)
21. Tibshirani, R.: Regression Shrinkage and Selection via the Lasso. *Journal of the Royal Statistical Society, Series B* 58(1), 267–288 (1996)
22. Xie, J., Zhang, L., You, J., Zhang, D.: Texture Classification via Patch-Based Sparse Texton Learning. In: *Int'l Conf. on Image Processing (ICIP)*, Hong Kong, pp. 2737–2740 (2010)

Selecting Anchor Points for 2D Skeletonization

Luca Serino and Gabriella Sanniti di Baja

Institute of Cybernetics “E. Caianiello”, CNR
Via Campi Flegrei 34, 80078 Pozzuoli, Naples, Italy
{l.serino,g.sannitidibaja}@cib.na.cnr.it

Abstract. In this paper two criteria are presented to compute reduced sets of centers of maximal discs in the weighted $\langle 3,4 \rangle$ distance transform of 2D digital patterns. The centers of maximal discs selected by the above criteria are used as anchor points in the framework of 2D skeletonization and, depending on the adopted criterion, originate skeletons with different properties.

1 Introduction

Skeletonization is a process leading to the extraction of a linear subset of a digital pattern, spatially placed along the medial region of the pattern and characterized by the same topology. The resulting set, the skeleton, is a stick-like representation of the pattern and, depending on the adopted process, accounts for different shape properties, such as symmetry, elongation, width and contour curvature.

A rich literature exists as concerns definition, extraction and use of the skeleton of 2D patterns (see e.g., [1-2] for an extensive survey). In particular, many algorithms deal with the computation of the *labeled skeleton*, i.e., a skeleton where each pixel is assigned the value of its distance from the complement of the pattern. The labeled skeleton can be used for shape representation even in the case of patterns that can not be interpreted as consisting exclusively of ribbon-like parts.

Most of the skeletonization algorithms, especially if dealing with the computation of the labeled skeleton, have been influenced by the work of Blum on the medial axis transform MAT [3]. Since the exact computation of the MAT is rather complex, different criteria have been suggested to generate an approximation of the MAT. One of these criteria is based on the detection of the centers of maximal discs CMD in the distance transform DT of the pattern. In fact, the CMD can be easily detected by comparing the distance values of neighboring pixels in DT, which is equivalent to comparing the radii of the discs associated to the neighboring pixels. Moreover, CMD are definitely symmetry points since their associated discs result to be tangent to the boundary of the pattern in at least two different boundary parts. Finally, each maximal disc is not included in any other single disc in the pattern and the union of the maximal discs coincides with the pattern. Unfortunately, the set of centers of maximal discs and the pattern are not generally characterized by the same topology. Thus, to obtain a topologically correct skeleton, also other pixels in DT have to be taken as skeletal pixels besides the CMD. To this aim, several DT based skeletonization algorithms have been suggested (see e.g., [4-12]).

DT based skeletonization algorithms can be implemented by following any of two different approaches. In both cases, CMD are taken as anchor points, i.e., as pixels that are definitely accepted as skeletal pixels. According to the first approach, the skeleton is computed by resorting to iterated contour peeling. At each iteration of the peeling process, pixels belonging to the current contour of the pattern and that are not anchor points are successively examined. Any such a pixel is removed, i.e., is assigned to the background, if its removal does not alter topology. Contour detection and peeling are iterated as far as pixel removal can be accomplished. DT is used not only for CMD detection, but also to reduce the computational cost of iterated peeling. In fact, repeated inspections of the image to identify the current contour of the pattern are avoided by processing for removal pixels of DT in increasing distance value order. According to the second approach, the skeleton is obtained by directly identifying in DT all the pixels constituting the skeleton. Once the anchor points have been detected, paths are grown in the direction of the increasing gradient from the anchor points having neighbors with larger distance value. Pixels detected by path growing are added to the set of the anchor points and path growing continues from them. This process originates a topologically correct skeleton in a fixed and small number of raster scan inspections of the image, independently of pattern's thickness. Whichever of the above two approaches is followed, the CMD play a crucial role to guarantee that the skeleton is symmetrically placed within the pattern and is characterized by the *recovery property*, i.e., the pattern can be faithfully reconstructed by the envelope of the discs associated to the skeletal pixels.

An important issue in DT based skeletonization is that the CMD of digital patterns are often very many, especially for natural shapes. Thus, accepting all the CMD as anchor points may originate skeletons with a too large number of non-significant branches, whose removal requires an elaborate post-processing pruning phase. Therefore, it is of interest to devise suitable criteria to filter out some, less significant, CMD and select as anchor points only a suitable subset of the set of CMD, so as to compute a skeleton more manageable and still adequate to represent the pattern.

In this paper, based on the experience that we have recently gained when working with 3D object skeletonization [13, 14], we suggest two criteria for the selection of reduced sets of CMD in the $\langle 3,4 \rangle$ weighted distance transform of 2D patterns, and use the selected CMD as anchor points for skeletonization. We show that CMD selected according to the first criterion are adequate to compute manageable skeletons from which the input patterns can be almost completely reconstructed. In turn, skeletons computed by selecting the anchor points by means of the second criterion are of interest in the framework of skeletonization at different levels of detail.

2 Preliminaries

We deal with binary images, where the pattern P is the set of 1's and the background B is the set of 0's, and use the 8-connectedness and the 4-connectedness for P and B , respectively.

The neighborhood $N(p)$ of a pixel p includes the eight pixels $n_i, i=1,2,\dots,8$, as shown in Fig. 1. The four n_i, i odd, and the four n_i, i even, are also termed edge-neighbors and vertex-neighbors of p , respectively.

n_2	n_3	n_4
n_1	p	n_5
n_8	n_7	n_6

Fig. 1. The eight neighbors of a pixel p

A *path* linking two pixels p and q is a sequence $p=p_0, p_1, \dots, p_s=q$, where p_i is a neighbor of p_{i-1} , for $1 \leq i \leq s$.

The *distance* between two pixels p and q is the length of a shortest path linking p to q [15,16]. The $\langle 3,4 \rangle$ weighted distance is obtained if the length is measured by respectively weighting 3 and 4 the unit moves towards edge- and vertex-neighbors encountered along the path [17]. The $\langle 3,4 \rangle$ weighted distance combines the simplicity of path-based distances with a reasonable approximation to the Euclidean distance.

The *distance transform* DT of P is a labeled replica of P, where the pixels of P are labeled with the length of a shortest path, entirely consisting of pixels of P, to the background B. In the following, we will refer to DT as to the distance transform computed by using the $\langle 3,4 \rangle$ weighted distance. We will denote the distance value of a pixel p in DT by $d(p)$. DT can be conveniently computed in two raster scans of the image, during which local distance information is sequentially propagated to the currently inspected pixel from its already visited neighbors [17].

Any pixel p in DT can be interpreted as the center of a disc with radius $d(p)$ included in the pattern. The disc associated to p can be constructed by applying to p the reverse distance transformation [17]. A pixel is a *center of maximal disc* CMD if the associated disc is maximal, i.e., is included in the pattern, but is not included by any other single disc in the pattern.

A pixel p of P is *simple* if its removal from P does not alter the topology of P. In other words, neither the number of 4-connected components of background pixels, nor the number of 8-connected components of pattern pixels should change in the 3×3 neighborhood of p by removing p . We count the number of 8-connected components of pattern pixels by means of the connectivity number $C_8(p)$ [18]. We also count the number $b(p)$ of background edge-neighbors of p . A pixel p is simple if the following condition is verified:

$$b(p) > 0 \text{ and } C_8(p) = 1$$

Simple pixels can be identified in DT by performing a suitable binarization of the distance values in $N(p)$ in order to distinguish the neighbors of p into pixels belonging to P and B.

The *skeleton* S of P is a subset of P with the following features: 1) it has the same topology as P; 2) it is symmetrically placed within P; 3) it consists of arcs and curves; and 4) its pixels are labeled with their distance from B.

As for the *recovery property* of the skeleton, is well known that if S includes all the CMD, then P can be completely reconstructed by the union of the discs centered on the pixels of S. It is also well known that it is seldom possible to obtain a skeleton that includes all CMD and is unit wide. Thus, if a unit wide skeleton is desired, some

pattern pixels may not be reconstructed from the skeleton. Moreover, also other pattern pixels may not be reconstructed if the skeleton undergoes pruning. On the other hand, pruning is often necessary to remove scarcely significant branches whose presence would otherwise make the structure of S too complex for a profitable use of the skeleton. Thus, in general, the recovery property is only partially satisfied by S . The ratio between the number of pixels actually recovered by the skeleton and the number of pixels constituting the input pattern can be used to measure the reconstruction ability of the skeleton. If the ratio is equal to 1, P is fully recovered from S .

3 Selecting the CMD in the Distance Transform

The distance transform can be interpreted as the result of a process during which a wavefront, originated at the boundary of a pattern P , propagates distance information with constant velocity towards the interior of P . Pixels that are reached at the same instant of time by the wavefront have the same distance value.

Centers of maximal discs are symmetrically placed within P . Thus, they are pixels of DT where the propagating wavefront (partially) folds upon itself. From these pixels, distance information is not propagated towards the interior of the pattern. Accordingly, CMD detection can be accomplished by comparing $d(p)$ with the distance values of the neighbors of p , by taking into account the weights 3 and 4 [19].

In detail, p is a center of maximal disc if for all its neighbors n_i , $i=1,2,\dots,8$, it results:

$$\begin{aligned} d(n_i) - d(p) &< 3, \text{ i odd} \\ \text{and} \\ d(n_i) - d(p) &< 4, \text{ i even} \end{aligned}$$

For completeness, we point out that the distance value 3 (6) must be replaced by the equivalent value 1 (5) before detecting the CMD. In fact, the discs with radii 3 (6) and 1 (5) are identical and replacement of 3 (6) by 1 (5) allows us to avoid erroneously detecting as CMD a pixel whose associated disc is not maximal [19].

Obviously, at least one neighbor of p in DT has distance value smaller than $d(p)$. In fact, the distance transform can be interpreted as due to a local propagation process during which any pixel p of P receives distance information from some of its neighbors, and propagates distance information to its neighbors in the pattern that are farther than p from B . Thus, at least one edge-neighbor of p is labeled $d(p)-3$, or at least one vertex-neighbor is labeled $d(p)-4$. As concerns the remaining neighbors of a CMD p , we note that some of them may have distance values larger than $d(p)$. In fact, edge-neighbors with value up to $d(p)+2$ and vertex-neighbors with value up to $d(p)+3$ do not prevent p from satisfying the above CMD condition.

If p is a CMD and the difference between $d(n_i)$ and $d(p)$ is the largest possible (i.e., it is 2 for edge-neighbors of p and 3 for vertex-neighbors), the importance of p is small. In fact, only a little part of the disc associated to p is not included in the disc associated to any of its neighbors with larger value. In turn, when the difference between $d(n_i)$ and $d(p)$ is smaller, the importance of p increases, since a larger part of

the disc associated to p is not included in the disc associated to one of its neighbors with larger value. See Fig. 2, where the discs associated to p and to its edge-neighbor n_7 are shown in the two cases ($d(p)=25, d(n_7)= 27$) and ($d(p)=25, d(n_7)= 26$).

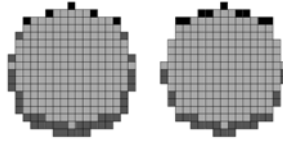


Fig. 2. Pixels belonging exclusively to the disc associated to p , with $d(p)=25$, and to the disc associated to n_7 , with $d(n_7)= 27$, left, and with $d(n_7)= 26$, right, are shown in black and in dark gray, respectively. Pixels belonging to both discs are in light gray.

The above considerations allow us to introduce a criterion to select a particular type of CMD that we call *relevant centers of maximal discs* RCMD. We say that a CMD p is a RCMD if for all its neighbors $n_i, i=1,2,\dots,8$, the following condition is satisfied:

$$\begin{aligned}
 & d(n_i) - d(p) < 2, \text{ i odd} \\
 & \text{and} \\
 & d(n_i) - d(p) < 3, \text{ i even}
 \end{aligned}$$

The second criterion for CMD selection is based on the notion of convexity. The *contour* of P consists of the pixels of P having at least one edge-neighbor in B . A contour pixel p is placed on a locally linear part of the contour of P if $N(p)$ includes three pixels belonging to the background B . In turn, p is placed in a *local convexity* of the contour if $N(p)$ includes more than three pixels of B . The larger the number of background neighbors of p is, the sharper the convexity in p is.

We extend the above notion on convexity, defined for the contour pixels, to the pixels in DT. In this case, for a pixel p with distance value $d(p)$, we measure the degree of convexity by counting the neighbors of p whose distance value is smaller than $d(p)$. In fact, pixels with the same distance value in DT are reached simultaneously by the propagating wavefront. In turn, pixels with smaller distance value have been reached by the propagating wavefront at previous instant of times, i.e., they are the pixels providing distance information to pixels with distance value $d(p)$ and can be, accordingly, interpreted as background pixels. Hence, we can characterize the pixels on each wavefront (and in particular the CMD in the wavefront) with their corresponding convexity degree.

Let $m(p)$ be the number of neighbors of a CMD p with distance value smaller than $d(p)$ and let θ be a threshold, whose value ranges from four to seven. Then, we say that a CMD p is a θ -convexity center of maximal ball θ -CMD if the following condition is satisfied:

$$m(p) \geq \theta$$

Obviously, the θ -CMD selected for a given θ constitute a proper subset of each of the sets of θ -convexity centers of maximal balls selected with a smaller value of θ .

4 RCMD as Anchor Points for Skeletonization

The CMD in the distance transform of a pattern P are generally not all equally important in the framework of skeletonization. For example, some CMD can be due only to the discrete nature of the digital space, as it is in the case of a digital circle. The digital counterpart of a continuous circle is unavoidably delimited by a polygonal line. Thus, weak convexities along the polygonal line may cause folding of the propagating wavefront upon itself, where CMD are detected. See Fig. 3.

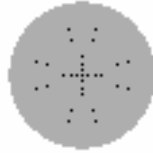


Fig. 3. A digital circle. The CMD are shown in black

For the continuous circle only one symmetry point exists, namely the center of the circle, and the continuous skeleton coincides with such a unique symmetry point. In the discrete case, if all CMD are taken as anchor points, the skeleton of a circle will not be just the center of the circle. On the other hand, if only the CMD in correspondence with the center of the circle (i.e., the most internal CMD) is selected as anchor point, the pattern reconstructed by applying to the skeleton the reverse distance transformation would remarkably differ from the digital circle.

More in general, if all CMD are taken as anchor points, a skeleton with a large number of scarcely significant branches is obtained. In turn, by selecting only the most internal CMD, found in correspondence with connected components of the wavefront entirely consisting of CMD, the skeleton would have a limited recovery property. Thus a compromise between the simplicity of the skeleton structure and a satisfactory pattern recovery is necessary. We suggest to use as anchor points only the RCMD, in order to obtain a skeleton having simple structure and still a good reconstruction ability.



Fig. 4. Test images

We accomplished skeletonization based on the selection of RCMD as anchor points on a number of binary images taken from a large publicly available dataset [20]. A set of ten test images taken from [20] is shown in Fig. 4. We follow the iterated peeling approach. Once relevant CMD have been marked as anchor points, we access pixels of DT in increasing distance value order. Pixels with the same distance value are sequentially removed, i.e., are assigned the background value 0, if they are not anchor points and are simple. As for the binarization of $N(p)$ necessary to

check whether p is simple, neighbors of p with value 0, or with value smaller than $d(p)$ and that are not marked as anchor points are interpreted as background pixels. All other neighbors of p are interpreted as pattern pixels.

Since RCMD may form a 2-pixel wide set, the skeleton is likely to be 2-pixel wide and its reduction to unit thickness is obtained by means of final thinning. To avoid unwanted shortening of skeleton branches, we perform final thinning by removing pixels for which at least one edge-neighbor is a background pixel and any of the templates in Fig. 5 is matched, where letters b and p denote background pixels and pattern pixels, respectively.

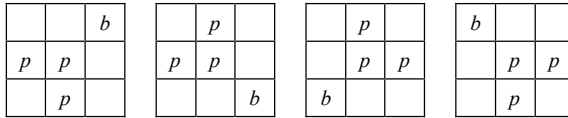


Fig. 5. Templates for final thinning

The skeletons of the ten test images obtained by taking as anchor points the RCMD can be seen in Fig. 6. We point out that pruning has not been performed. Thus, a few scarcely significant branches may still exist in the skeletons.

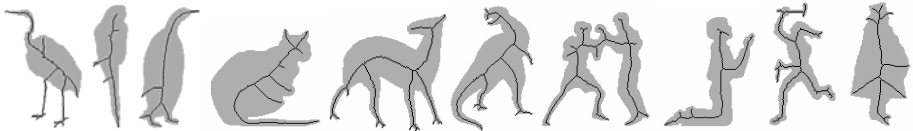


Fig. 6. Skeletons computed by using the RCMD as anchor points

The skeletons in Fig. 6 can be compared with the skeletons in Fig. 7, which are obtained by using the same algorithm, but by taking as anchor points all the CMD instead of only the RCMD. It is evident that the skeletons computed when only RCMD are selected are characterized by noticeably simpler structure.



Fig. 7. Skeletons computed by using all the CMD as anchor points

The performance of skeletonization based on the selection as anchor points of the RCMD can be quantitatively appreciated in Table 1. The first (second) row orderly shows for the ten test patterns the reconstruction ability of the skeleton based on the RCMD (on the CMD) measured as the ratio between the number of recovered pixels and the number of pixels in the input pattern. We note that the reconstruction ability of the skeleton based on RCMD is equal to or at most only slightly smaller than the

reconstruction ability of the skeleton computed by taking all CMD as anchor points. At the same time, the comparison between Fig. 6 and Fig. 7 shows that filtering out the centers of maximal discs that are not RCMD allows us to simplify the structure of the skeleton.

Table 1. Reconstruction ability of the skeleton

Skeleton based on RCMD	0.98	0.98	0.99	0.97	0.99	0.99	0.98	0.97	0.97	0.99
Skeleton based on CMD	0.98	0.99	0.99	0.99	0.99	0.99	0.99	0.99	0.98	0.99

5 θ -CMD as Anchor Points for Skeletonization

Though the recovery property is important to guarantee that the skeleton is a faithful representation of a pattern, in some cases a rougher representation may still be enough to give a sketched version of the shape of the pattern. Thus, it is of interest a criterion to select different subsets of the CMD able to originate different skeletons, from the more detailed skeletons to the less detailed ones.

The notion of convexity introduced in Section 3 can be used to select subsets of the CMD based on their convexity degree and to originate skeletons with different levels of detail. As an example, refer to Fig. 8, showing the three sets of θ -convexity centers of maximal balls that have been selected for one of the test images by using three different values of the threshold on the number $m(p)$ of neighbors of a CMD p with distance value smaller than $d(p)$. Also the three corresponding skeletons are shown in Fig. 8. The same skeletonization algorithm described in the previous Section is used also in this case, but the anchor points are now the θ -CMD. As before, final thinning has been employed to obtain a unit wide skeleton and pruning has not been accomplished.

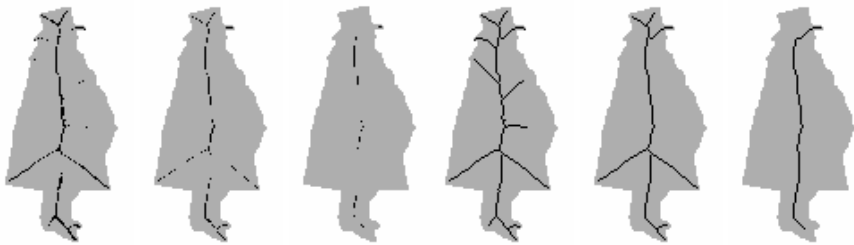


Fig. 8. From left to right, the three sets of θ -CMD with $\theta=5$, $\theta=6$, and $\theta=7$, and the three corresponding skeletons

It can be seen that the number of peripheral branches of the skeleton diminishes when θ increases. Of course, while the structure of the skeleton becomes simpler, the representative power of the skeleton, in terms of reconstruction ability, diminishes. For large values of θ , the skeleton represents a sketched version of the input pattern, where some details are missing. Refer to Fig. 9, where the patterns reconstructed by

applying the reverse distance transformation to three skeletons based on the three sets of θ -CMD for $\theta=5$, $\theta=6$, and $\theta=7$ are shown from left to right. In the three cases, the reconstruction ability of the skeleton is respectively 0.99, 0.98 and 0.90.



Fig. 9. From left to right, the patterns recovered by applying the reverse distance transformation to the skeletons based on θ -CMD with $\theta=5$, $\theta=6$, and $\theta=7$. Pattern's pixels that are not recovered are shown in black.

6 Conclusion

In this paper we presented two criteria to compute reduced sets of centers of maximal discs in the weighted $\langle 3,4 \rangle$ distance transform of 2D digital patterns and used the selected centers of maximal discs as anchor points for skeletonization.

In the literature, all the centers of maximal discs have been generally used as anchor points in the framework of skeletonization, but this leads to skeletons with a complex structure where not all branches are actually significant. We here suggested a criterion to select a subset of the CMD, namely the relevant centers of maximal balls RCMD that, used as anchor points, allow us to obtain skeletons with simple structure and still characterized by a satisfactory reconstruction ability. We also suggested a second CMD selection criterion, based on the notion of convexity, which identifies the θ -convexity centers of maximal balls θ -CMD and is useful in the framework of skeletonization at different levels of detail. Only CMD on local convexities of the wavefronts of DT characterized by a degree of convexity larger than a threshold are selected as anchor points. By changing the value of the threshold θ , different skeletons are obtained.

The characterization of the CMD has been discussed with reference to the $\langle 3,4 \rangle$ weighted distance transform. However, we believe that this characterization can be simply extended to distance transforms computed by using a larger neighborhood (e.g., the 5×5 neighborhood, where also the knight move is taken into account).

The algorithm has been implemented in C and runs on a Pentium 4 (3 GHz, 2 GB RAM) personal computer. It has been tested on about 100 patterns with different shape and size taken from [20].

References

- [1] Lam, L., Lee, S.W., Suen, C.Y.: Thinning methodologies-a comprehensive survey. *IEEE Trans. PAMI* 14(9), 869–885 (1992)
- [2] Saeed, K., Tabezki, M., Rybnik, M., Adamski, M.: K3M: a universal algorithm for image skeletonization and a review of thinning techniques. *Int. J. Appl. Math. Comput. Sci.* 20(2), 317–335 (2010)
- [3] Blum, H.: A transformation for extracting new descriptors of shape. In: Wathen-Dunn, W. (ed.) *Models for the Perception of Speech and Visual Form*, pp. 362–380. MIT Press, Cambridge (1967)
- [4] Arcelli, C., Sanniti di Baja, G.: A width-independent fast thinning algorithm. *IEEE Trans. PAMI* 7, 463–474 (1985)
- [5] Klein, F.: Euclidean skeletons. In: *Proc. 5th Scand. Conf. Image Anal.*, pp. 443–450 (1987)
- [6] Arcelli, C., Sanniti di Baja, G.: A one-pass two-operations process to detect the skeletal pixels on the 4-distance transform. *IEEE Trans. PAMI* 11, 411–414 (1989)
- [7] Xia, Y.: Skeletonization via the realization of the fire front's propagation and extinction in digital binary shapes. *IEEE Trans. PAMI* 11(10), 1076–1086 (1989)
- [8] Arcelli, C., Sanniti di Baja, G.: Euclidean skeleton via center-of-maximal-disc extraction. *Image and Vision Computing* 11, 163–173 (1993)
- [9] Kimmel, R., Shaked, D., Kiryati, N.: Skeletonization via distance maps and level sets. *Computer Vision and Image Understanding* 62(3), 382–391 (1995)
- [10] Sanniti di Baja, G., Thiel, E.: Skeletonization algorithm running on path-based distance maps. *Image and Vision Computing* 14, 47–57 (1996)
- [11] Svensson, S., Borgefors, G., Nystrom, I.: On reversible skeletonization using anchor-points from distance transforms. *Journal of Visual Communication and Image Representation* 10, 379–397 (1999)
- [12] Mekada, Y., Toriwaki, J.I.: Anchor point thinning using a skeleton based on the Euclidean distance transformation. In: *Proceedings of ICPR 2002*, vol. 3, pp. 923–926 (2002)
- [13] Serino, L., Arcelli, C., Sanniti di Baja, G.: A Characterization of Centers of Maximal Balls in the $\langle 3,4,5 \rangle$ weighted distance transform of 3D digital objects. In: *Proceedings WADGMM*, pp. 12–16 (2010)
- [14] Arcelli, C., Sanniti di Baja, G., Serino, L.: Distance driven skeletonization in voxel images. *IEEE Trans. PAMI* (2010), <http://doi.ieeecomputersociety.org/10.1109/TPAMI.2010.140>
- [15] Yamashita, M., Ibaraki, T.: Distances defined by neighborhood sequences. *Pattern Recognition* 19(3), 237–246 (1986)
- [16] Das, P.P., Chakrabarti, P.P., Chatterji, B.N.: Distance functions in digital geometry. *Information Sciences* 42, 113–136 (1987)
- [17] Borgefors, G.: Distance transformations in digital images. *Computer Vision, Graphics, and Image Processing* 34(3), 344–371 (1986)
- [18] Yokoi, S., Toriwaki, J.I., Fukumura, T.: An analysis of topological properties of digitized binary pictures using local features. *Comp. Graphics and Image Processing* 4, 63–73 (1975)
- [19] Arcelli, C., Sanniti di Baja, G.: Weighted distance transforms: a characterization. In: *Cantoni, V., Di Gesù, V., Levioldi, S. (eds.) Image Analysis and Processing II*, pp. 205–211. Plenum Press, New York (1988)
- [20] <http://www.lcms.brown.edu/~dmc/>

Interactive Segmentation of 3D Images Using a Region Adjacency Graph Representation

Ludovic Paulhac, Jean-Yves Ramel, and Tom Renard

Université François Rabelais Tours, Laboratoire Informatique (EA2101)
{ludovic.paulhac, jean-yves.ramel, tom.renard}@univ-tours.fr

Abstract. This paper presents an interactive method for 3D images segmentation. This method is based on a region adjacency graph representation that improves and simplifies the segmentation process. This graph representation allows the user to easily define some splitting and merging operations which gives the possibility to make an incremental construction of the final segmentation. To validate the interest of the proposed method, our interactive proposition has been integrated into a volumetric texture segmentation process. The obtained results are very satisfactory even in the case of complex volumetric textures. This same system, including the textural features and our interactive proposition, has been manipulated by specialists in sonography to segment 3D ultrasound images of the skin. Some examples of segmentation are presented to illustrate the interactivity of our approach.

Keywords: Interactive segmentation, 3D images, Graph.

1 Introduction

Image segmentation is an important topic in image analysis and computer vision and concerns many domains of application. The purpose of the segmentation is to partition images into regions that are in some sense homogeneous, or to isolate from the background one or several objects of interest. It is then possible to exploit the results of the segmentation to compute characteristics corresponding to isolated objects, to produce some visualizations, to follow an object in a video, etc.

To resolve various segmentation problems, numerous automatic methods have been proposed in the literature [1,2]. Nevertheless, it is sometimes difficult to obtain the desired results using an automatic process. The quality of an image acquisition or the abnormalities in a scene can lead to some variations which can increase the difficulties of building robust static segmentation methods. Moreover, automatic methods are often dedicated to specific problems and do not have a general applicability.

Interactive segmentation algorithms provide a solution to these problems. Interaction should allow the operator to drive and improve segmentation computation according to the kind of the images being processed. A great number of interactive methods have been developed mainly for 2D images [3,4]. Usually, these methods are proposed to process medical images. Among the interactive volume segmentation methods (volume segmentation systems), we can find the systems proposed in [5,6,7,8,9,10]. As in 2D,

they are usually dedicated to medical image segmentation and none of them proposes a similar representation as the one described below.

This paper describes in a formal way how it is possible to design a powerful and generic segmentation system based on a Region Adjacency Graph (RAG). Using the RAG representation, our purpose is to allow an operator to construct a segmentation progressively using split and merge operations. A similar approach included in a framework has been presented in [11]. Nevertheless, this method has not been proposed for an interactive purpose. Moreover, it is based on a 2D oriented boundary graph that limits the interaction with the user in a 3D domain.

In section 2, the design of the proposed segmentation system is presented using a general view of an interactive process. Section 3 describes the region adjacency graph content and construction. Section 4 presents the possible operations defined in order to allow the user to incrementally transform the RAG and in the same way improve the segmentation. In section 5, a concrete implementation of our model is realized to create an interactive system for volumetric texture segmentation. Then, the system is evaluated (considering the benefits of its interactivity or not). In section 6, this same system, including textual features and our interactive proposition, is used by specialists in sonography to segment 3D ultrasound images. To conclude, we provide a discussion about our work and introduce the main prospects.

2 Design of an Interactive Segmentation System

Figure 1 presents a general interactive system for image segmentation. As explained in [3], such a system contains different main components including the computational and the interactive parts.

The computational part corresponds to a set of methods capable to generate a segmentation. To be more precise, numerical features (F) computed by Features Extractors are exploited by one or several Segmentation Processes to generate a segmentation. In a 3D segmentation system, the set of Features F is computed for each used voxels in the image. Generally, all these methods use some parameters determined by prior knowledge or defined by the user in the interactive part. In Figure 1, the Feature Extraction Configuration component and the Segmentation Parameters component allow respectively the user to tune the feature extraction and the segmentation process.

Putting the user in the loop (interactive system) means proposing an incremental segmentation process to the user. Then, two problems have to be solved during the conception of the system. How to let the user defining the criteria to use during the segmentation computation ? Which feedback (representation of the results) and interactive tools (operators) to propose to the user to make this representation evolving in the desired direction ?

To answer the first question, we think that the best choice is to use a region based approach (clustering method) using criteria (features) defined by the user. The second problem is solved by structuring the huge amount of data to process with a region adjacency graph (RAG). This data structure constructed using the Computational part information (section 3) allows us to define the user interaction. By coupling this component with the Split/Merge Interactive Operations component and by analyzing the

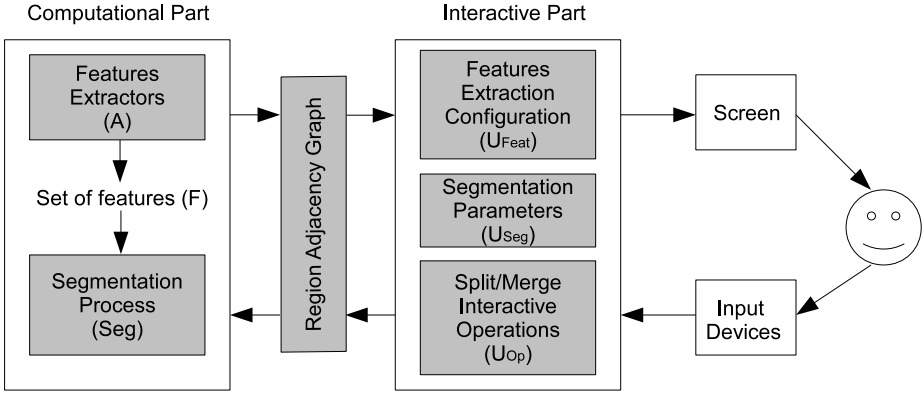


Fig. 1. The main components of our volume segmentation system

visual information of the segmentation given by the RAG (feedback), the user is able to define some actions to drive the segmentation process in an incremental way. In other words, the Split and Merge operations allow the user to manage the Region Adjacency Graph evolution (section 4).

3 Construction of the RAG

3.1 Selection of the Features to be Used during the Incremental Segmentation

We define as A the set of possible image features extractors. Each of them is able to provide one or more image features F . We can then write the following belonging relation:

$$a_{i,j} \in A \text{ with } i \in \{1..N\} \text{ and } j \in \{1..M_i\}$$

with $a_{i,j}$ the extractor j that allows to compute the feature i , N the number of possible image features and M_i the number of extractors allowing to obtain the feature i .

Image features can then be computed as follows:

$$a_{i,j} : D, U_{Feat_j} \mapsto F_i$$

with D the set of voxels to process, U_{Feat_j} the parameters of the feature computation method j that the interactive part should allow to tune easily.

Among the available features, the user chose the features $F = \{F_1, F_2, \dots\}$ using the interactive part. They are computed for each voxel of the image by running the corresponding algorithm in the computational part ($a_{i,j}$).

3.2 RAG Definition

We define $G(V, E)$ the region adjacency graph with V the set of vertices and E the set of edges. At the beginning of the graph representation, the number of vertices in

the graph is 1 (number of edges $E = 0$). Indeed, the image to be processed is considered as a single region represented by one vertex. As we will see in section 4, some operations associated to the graph have been defined in order to increase or reduce the number of vertices (then the number of regions) to produce the desired segmentation. In consequence, a vertex V of the graph represents a region of the image, and for each vertex we associate the following information: the average features \bar{F} of the region that corresponds to the centroid of the features F in the region, the center of gravity \bar{G} of the region that corresponds to the position of the vertex.

To connect the vertices (edge constructions), it is necessary to identify the adjacent regions for each of them. To do so, the intersection (a surface) between each region is computed. If its surface is not null then the two corresponding vertices are connected. Each edge E contains the information of the two linked vertices and the intersection of the two regions.

4 Interactive Segmentation Scheme

As we have seen previously, the RAG is initialized with a single vertex which is associated to the image to be processed. In consequence, the first step in the process of segmentation is a splitting operation (sub-section 4.1). In the following, the operator can define some parameters, some actions of splitting or merging in order to provide an evolution of the segmentation. The segmentation progression can be written as follows:

$$Seg : F, U_{Seg}, G_k, U_{op} \mapsto G_{k+1}$$

For each iteration, the user has the choice to define U_{op} where the splitting and merging operations are specified. Using U_{op} , the graph G_k , features F and the segmentation parameters U_{Seg} , the function Seg is able to generate a new segmentation, a new region adjacency graph G_{k+1} . According to the chosen action in U_{op} , the function Seg transforms the RAG representation (the segmentation) using formal operators described in subsection 4.1 and 4.2. Then, the user can decide to stop this incremental process when the desired segmentation is obtained.

In the previous section, the proposed formal representation based on the region adjacency graph has been presented. In the following sub-sections, we detail how the splitting and merging operations in our system work.

4.1 Splitting

If the user chooses to split a node ($U_{op}.action = split$), then the function Seg corresponds to a K-means clustering. We chose to use this method because the processing time of 3D images is sometimes huge and the main advantages of the K-means method are its speed and its low memory cost. Moreover it allows an efficient clustering of voxels.

The new graph representation G_{k+1} is obtained as follows:

$$G_{k+1} = Kmeans(U_{op}.V, G_k, F, U_{Seg})$$

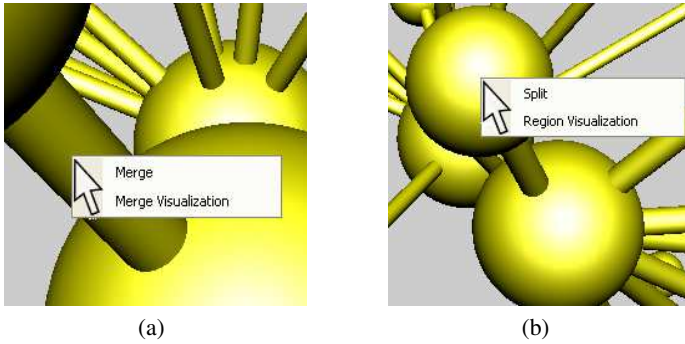


Fig. 2. Available actions with the RAG

with $U_{op}.V$ the vertex to which the splitting operation is applied, G_k the actual RAG and F the set of features. Here, U_{Seg} contains the specified number of classes K for the K-means function.

Using this clustering function, the corresponding region is divided into different areas. For each of them, a new vertex is created in the graph G_{k+1} . As explained in section 3 the average feature \bar{F} is initialized with the centroid of the features F in the region. The attribute \bar{G} is initialized with the center of gravity of the region. Moreover according to the adjacency surface between each new region, additional edges are added.

4.2 Merging

In this case ($U_{op}.action = merge$), the function Seg can be assimilated to a simple merging function $Merge$ where the two merged regions are identified by a new vertex V_{new} . The new graph representation G_{k+1} is then obtained as follows:

$$G_{k+1} = Merge(U_{op}.E, G_k)$$

with $U_{op}.E$ the edge that contains the two vertices to merge and G_k the actual RAG.

To compute the attributes $T = \{\bar{F}, \bar{G}\}$ of the new node V_{new} using the two vertices V_1 and V_2 , the following operation is applied:

$$V_{new}.T_i = \frac{(V_1.T_i)(V_1.NV) + (V_2.T_i)(V_2.NV)}{V_1.NV + V_2.NV}$$

where NV corresponds to the number of voxels identified by a node in the region.

4.3 Interactive Segmentation Using the RAG Visualization

To manipulate the segmentation the user uses the 3D region adjacency graph displayed in an OpenGL window (Figure 5 and 6). The operations of merging and splitting are available by clicking on the nodes and on the edges inside the OpenGL Window (Figure 2). The result of the chosen operation is then updated on the screen for the user

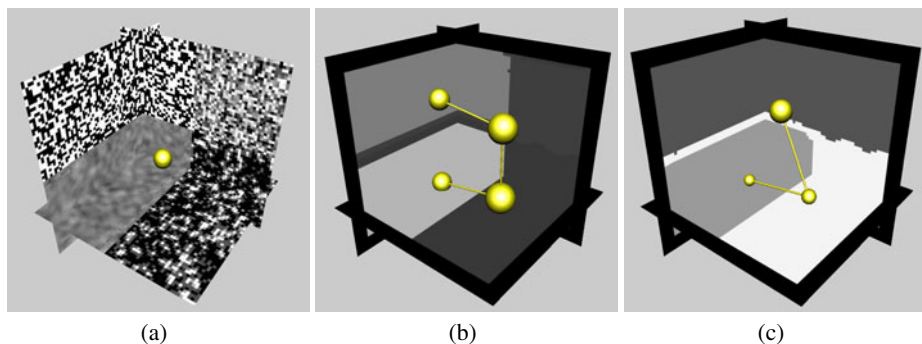


Fig. 3. Example of the RAG evolution

feedback. The region adjacency graph representation is also an asset in visualization. The user can easily visualize the structure of the image content and the different identified regions. In our representation, the size of the vertex depends on the number of voxels inside each of the segmented areas. This allows the user to identify the main regions, the main components in the processed image. To guide the user in its choices, an OpenGL visualization of the regions is available by clicking on the corresponding vertex. It is also possible to display a preview of the merge representation result by clicking on the corresponding edges. This allows the user to visualize what should be the new segmentation result with the merge operation (Figure 2).

Figure 3 shows an example of the region adjacency graph evolution. First the image that is considered as a single region is represented by one vertex (Image 3 (a)). Then, the user performs a splitting operation with $U_{Seg} = 4$ to identify the 4 regions. In the image 3 (b), each of them is represented by one node and two nodes are not linked because their corresponding regions are not identified as adjacent. At the end, the user choose to merge the two upper nodes using the merging operation (Image 3 (c)).

5 Evaluation of Our Interactive System in a Texture Segmentation Problem

In this part, our purpose is to prove the interest and the efficiency of our model. To do so, we compare 2 systems using the same texture features: the first one does not propose any interaction, the second one proposes to use the split and merge operations of our RAG representation.

To describe the voxels of 3D images, we choose the 3D texture features presented in [12], inspired by the human way to describe a texture. They are easily understandable by humans and, in the sense, these texture features make the setting up of our interactive system easier. Indeed, it is more convenient for a user to select the features which are interesting to use in a given application. The proposed characteristics (F_i) are: Granularity, which can be represented by the number of three-dimensional patterns constituting the texture, shape information about these patterns with the volume and the compactness, regularity of these patterns, directionality that measures the strong or weak presence of a privileged direction, contrast and roughness of the image, which are

also important information. During the segmentation process, the proposed features are computed for each voxel and for several resolutions. Then, a voxel of the initial image is described by a vector containing $7n$ different features with n the number of resolutions and 7 the number of proposed features. The K-means algorithm [13] allows to generate a segmentation using the set of computed vectors.

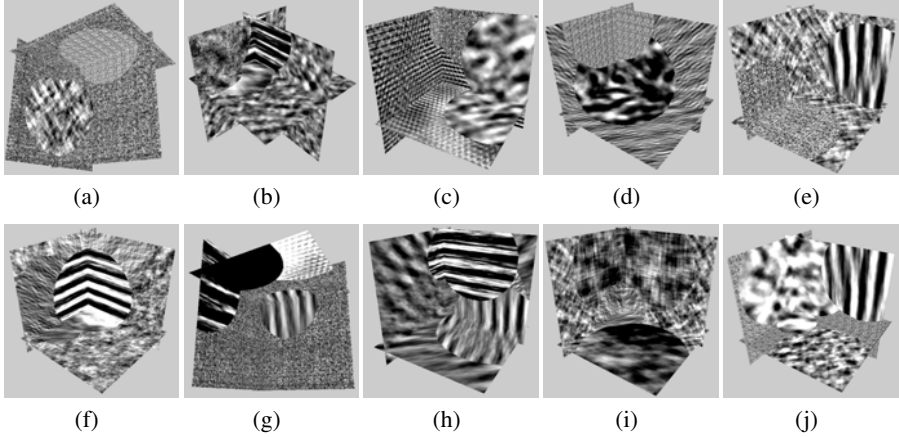


Fig. 4. Volumetric texture images : [a-e] 3 classes of textures, [f-j] 4 classes of textures

During our experiments, ten 3D textured images have been used (Figure 4): 5 images with 3 classes of textures and 5 images with 4 classes of textures. To generate segmentations, different cases have been considered. In the first one, the user works with a system without any interaction (WI). The number of the chosen classes (in U_{Seg}) for the K-means algorithm corresponds to the number of volumetric textures inside the processed 3D image. The second one uses the split and merge operations of our RAG representation that we described the section 4.

To give an evaluation of the produced segmentation, the generic discrepancy measure d_{gdm} [14] was used. This measure uses the computation of a distance between partitions that have been defined by Gusfield [15]. If d_{gdm} is equal to 0, then the segmentation is ideal whereas an inverse segmentation generates the value 1.

Table 1 shows the normalized partition distance for each solid texture segmentation. To have a better readability, results of the generic discrepancy measure have been multiplied by 100 and are comprised between 0 and 100.

As expected, the best results are obtained with the interactive system. Without any interaction, the system cannot manage the different image specificities and it is difficult to obtain good results of segmentation. In Table 1, the system without interaction gives the lowest results every time. Providing the system with only the number of classes inside the processed image is not sufficient. Sometimes, several regions can be identified inside a same texture. With the interactive system, the user can sometimes avoid this problem. By merging the different regions, he can reach the best segmentation. Indeed, it is possible to merge regions when different classes are identified inside a unique

texture, and it is possible to focus the system inside a region using the split operation when a texture has not been correctly identified in the current segmentation. For all the computed segmentations, the obtained results are very satisfactory even if the processed 3D images contain complex volumetric textures.

Table 1. Comparison of segmentation results using the generic discrepancy measure

Number of classes	Systems	Image (a) to (j)				
3 classes	WI	11.78	12.07	21.61	19.53	27.98
	RAG	2.94	2.80	2.46	3.12	5.59
4 classes	WI	25.96	14.87	29.56	44.37	28.91
	RAG	12.05	2.10	7.38	10.41	11.72

Of course, a fully automatic system is different from an interactive one and in some cases interactivity is not possible (when the amount of data to be processed is very huge or when there are time constraints). Our proposition does not deal with such applications but concerns the numerous applications in which an expert has to interpret the result of the segmentation process. We believe that showing just the final segmentation result (even depending of some basic settings) is frustrating for the experts. It is pleasant to participate to the segmentation process especially if the final results are better. This is the case for most medical imaging systems. Furthermore, providing a system that let the user act with complete freedom sometimes provides very interesting results obtained in a roundabout way. The next section demonstrates this kind of utilization on a real life application in the medical domain.

6 Segmentation of 3D Ultrasound Images

A software built using the concepts proposed in this paper has been manipulated by specialists in sonography to segment 3D ultrasound images of the skin. To illustrate the usability of this software, 2 scenarios of utilization are presented.

The first scenario allows the segmentation of a nevus presented in Image 5(a) that is a 3D ultrasound image of the skin. As it is possible to see in Image 5(b), just a part of the nevus has been identified by applying a first split operation on the initial RAG (one node = the image). Here the Seg function corresponds to a K-means algorithm. In this case, the user choses a number of classes ($U_{Seg} = 3$) that is too low to find the desired region. It is then necessary to focus on the blue region in Image 5(c). To do so, the user uses the region adjacency graph to select the corresponding region. It is then possible for him to make a visual representation of the region in Image 5(c) or to start a new segmentation of the selected region 5(d). After the splitting operation (that used $U_{Seg} = 2$) of the blue region Image 5(c), Image 5(e) has been obtained. The nevus is then identified but composed of two regions. By using the region adjacency graph, these two regions can be merged to obtain a final segmentation Image 5(f). The nevus is represented by one yellow area and it is then possible, for the user, to isolate it using a mesh visualization (Image 5(g)) or to compute different kinds of features like its volume, for example.

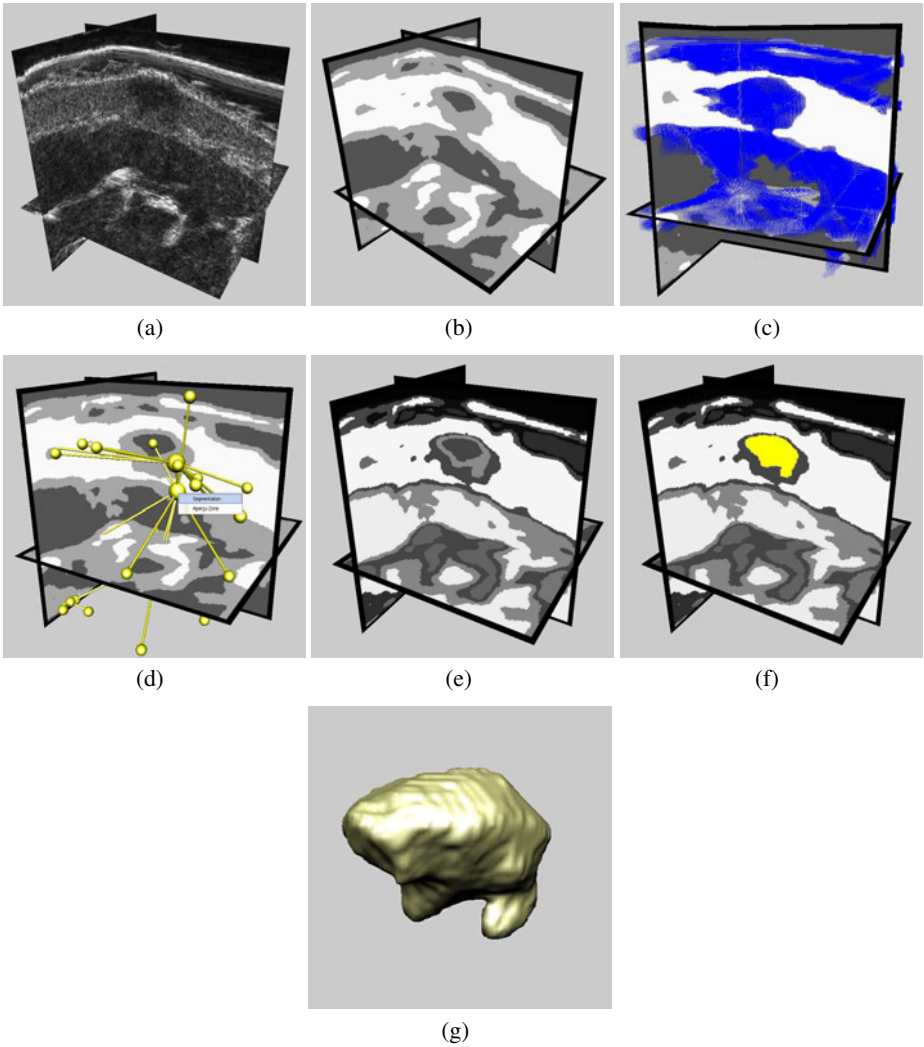


Fig. 5. Segmentation of a nevus using the RAG interactive method

Figure 6 presents the second scenario of segmentation. The processed image is a 3D ultrasound image of the skin that contains a tendon. Here, the aim is to produce a segmentation that allows the user to isolate the tendon area. The segmentation is realized in a zone of interest around the tendon defined by the purple box (Image 6(a)).

A first splitting operation is applied with $U_{Seg} = 8$ to generate the initial segmentation Image 6(b). Using merging operations, the user improves the results of segmentation by merging the regions of the tendon that are isolated in the segmentation Image 6(b). The final segmentation presented in Image 6(c) has been obtained and Image 6(d) shows its region adjacency graph representation. As previously, we show a

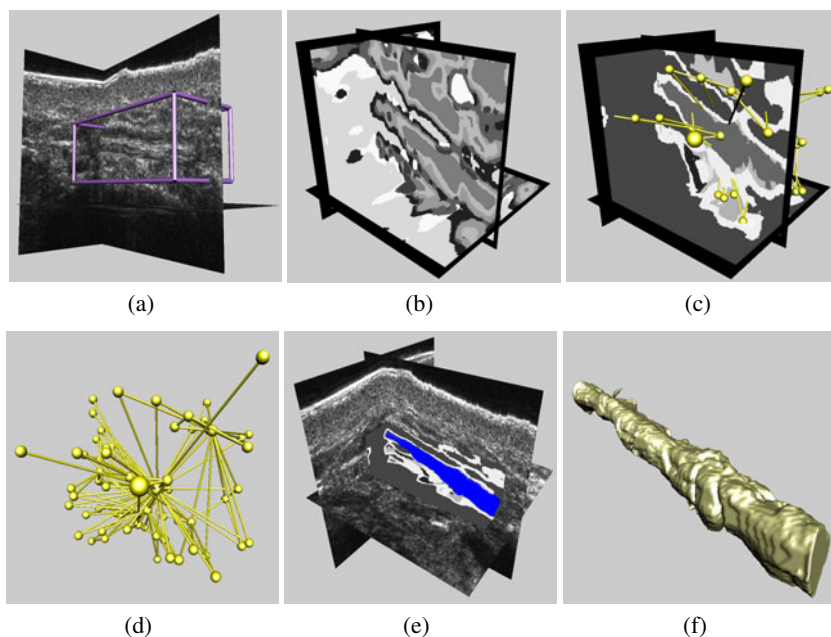


Fig. 6. Segmentation of a tendon using the RAG interactive method

representation of the final segmentation inside the initial image (Image 6(e)) and a mesh visualization of the tendon (Image 6(f)).

7 Conclusion

In this paper, a general scheme for interactive segmentation system conception has been presented. It is based on a region adjacency graph representation and a list of features associated to each voxel (that can be selected by the user at each step of the segmentation process). The vertices of the graph are positioned using the center of gravity of each region in the segmented image. Two nodes in the graph are connected if the adjacency surface between the corresponding region is not null. This graph allows a user to define merging and splitting operations to incrementally improve the segmentation results. To merge two regions, the user clicks on their corresponding edge, and to split a region the user can select features and parameters to use before running the merging operation by clicking on the corresponding node. To evaluate our proposition, a concrete case of texture segmentation has been presented. Using the proposed interactive method, the user greatly improves the results of segmentation. The same system using textural features has been used to segment 3D ultrasound images. Our software has been manipulated by specialists in sonography that appreciated the flexibility of our proposition. Whatever the situation, the user can improve the segmentation results using merging operations or region focusing.

Several improvements could be added in the future. As we can see in Image 6(c) the number of generated vertices is sometimes important. This can be a problem for

the user when he tries to find a particular region. Then it could be interesting to propose some automatic merging methods in order to reduce the graph complexity without losing useful information. Finally, the structural information contained in the region adjacency graph representation could be compared to an atlas in order to guide segmentation algorithms automatically.

References

1. Campadelli, P., Casiraghi, E., Exposito, A.: Liver segmentation from computed tomography scans: A survey and a new algorithm. *Artificial Intelligence in Medicine* 45, 185–196 (2009)
2. Oliver, A., Freixenet, J., Martí, J., Pérez, E., Pont, J., Denton, E.R., Zwiggelaar, R.: A review of automatic mass detection and segmentation in mammographic images. *Medical Image Analysis* 14, 87–110 (2010)
3. Olabarriaga, S., Smeulders, A.: Interaction in the segmentation of medical images: A survey. *Medical Image Analysis* 5, 127–142 (2001)
4. McGuinness, K., O'Connor, N.E.: A comparative evaluation of interactive segmentation algorithms. *Pattern Recognition* 43, 434–444 (2010)
5. Boykov, Y., Jolly, M.-P.: Interactive organ segmentation using graph cuts. In: Delp, S.L., DiGoia, A.M., Jaramaz, B. (eds.) *MICCAI 2000*. LNCS, vol. 1935, pp. 276–286. Springer, Heidelberg (2000)
6. Bartz, D., Mayer, D., Fischer, J., Ley, S., del Rio, A., Thust, S., Heussel, C., Kauczor, H.U., Strasser, W.: Hybrid segmentation and exploration of the human lungs. In: *VIS 2003: IEEE International Conference in Visualization*, pp. 177–184 (2003)
7. Gu, L., Peters, T.: Robust 3d organ segmentation using a fast hybrid algorithm. *Computer Assisted Radiology and Surgery* 1268, 69–74 (2004)
8. Tzeng, F.Y., Lum, E., Ma, K.L.: An intelligent system approach to higher-dimensional classification of volume data. *IEEE Transactions on Visualization and Computer Graphics* 11, 273–284 (2005)
9. Ben-Zadok, N., Riklin-Raviv, T., Kiryati, N.: Interactive level set segmentation for image-guided therapy. In: *ISBI 2009: IEEE International Symposium on Biomedical Imaging*, pp. 1079–1082 (2009)
10. Prabni, J.S., Ropinski, T., Hinrichs, K.: Uncertainty-aware guided volume segmentation. *IEEE Transactions on Visualization and Computer Graphics* 16, 1358–1365 (2010)
11. Baldacci, F., Braquelaire, A.J.P., Domenger, J.P.: Oriented boundary graph: A framework to design and implement 3d segmentation algorithms. In: *ICPR 2010: 20th International Conference on Pattern Recognition*, pp. 1116–1119 (2010)
12. Paulhac, L., Makris, P., Gregoire, J.M., Ramel, J.Y.: Human understandable features for segmentation of solid texture. In: *Bebis, G., Boyle, R., Parvin, B., Koracin, D., Kuno, Y., Wang, J., Wang, J.-X., Wang, J., Pajarola, R., Lindstrom, P., Hinkenjann, A., Encarnação, M.L., Silva, C.T., Coming, D. (eds.) ISVC 2009*. LNCS, vol. 5875, pp. 379–390. Springer, Heidelberg (2009)
13. Coleman, G., Andrews, H.: Image segmentation by clustering. *Proceedings of the IEEE*, 773–785 (1979)
14. Cardoso, J.S., Corte-Real, L.: Toward a generic evaluation of image segmentation. *IEEE Transactions on Image Processing* 14(11), 1773–1782 (2005)
15. Gusfield, D.: Partition-distance: A problem and class of perfect graphs arising in clustering. *Information Processing Letters* 82(9), 159–164 (2002)

An Algorithm to Detect the Weak-Symmetry of a Simple Polygon

Mahmoud Melkemi¹, Frédéric Cordier¹, and Nickolas S. Sapidis²

¹ Université Haute Alsace, Laboratoire de Mathématiques Informatique et Applications, 4 rue des frères Lumière, 68093, Mulhouse, FRANCE

`mahmoud.melkemi@uha.fr`, `frederic.cordier@uha.fr`

² Department of Mechanical Engineering, University of Western Macedonia
Bakola & Sialvera Str., Kozani GR-50100, GREECE

`nsapidis@uowm.gr`

Abstract. This article deals with the problem of detecting the weak-symmetry of a simple polygon. The main application of this work is the automatic reconstruction of 3D polygons (planar or non-planar polylines) symmetric with respect to a plane from free hand sketching 2D polygons. We propose a provable approach to check on the weak-symmetry of a simple polygon. The worst time complexity of the proposed algorithm is $O(n^3)$ where n is the number of the vertices of the input polygon.

1 Introduction

The 3D reconstruction from freehand sketches is an important problem in computer vision and computer graphics (see for example [1]). Given a set of 2D polygons provided by the user (drawn by the user on the plane $z = 0$), the 3D reconstruction consists of computing the 3D polygons (planar or non-planar polylines) such that their orthogonal projection matches the input 2D polygons. The x and y coordinates of the vertices of the reconstructed polygons are known. The z -coordinates have to be computed. The difficulty is that for each vertex of the 2D polygons, there exist an infinite number of 3D vertices whose orthogonal projection matches the 2D vertices. In this paper, we consider the reconstruction of mirror-symmetric 3D polygons (orthogonal symmetric with respect to a central plane) from their orthogonal projection, it involves two steps: (1) finding what are the pairs of vertices symmetric to each other, (2) using this correspondence called *weak-symmetry*, computing the vertex positions of the mirror-symmetric 3D polygon.

The notion of weak-symmetry we will study here comes from the following property: let V and V' be the two sets of vertices of a 3D polygon which are mirror-symmetric to each other. Let $V_p = \{v_{p,0}, \dots, v_{p,i}, \dots, v_{p,n-1}\}$ and $V'_p = \{v'_{p,0}, \dots, v'_{p,i}, \dots, v'_{p,n-1}\}$ be the orthogonal projections of the two sets V and V' ($v_{p,i}$ and $v'_{p,i}$ are respectively the projections of $v_i \in V$ and of its mirror-symmetric $v'_i \in V'$).

Property 1: *The straight lines that join $v_{p,i}$ and $v'_{p,i}$ are parallel to each other.* For more detail see [2]. The correspondence between these vertices is called the *weak-symmetry*.

Problem Statement. Given a 2D polygon which is an orthogonal projection of an unknown mirror-symmetric 3D polygon. This paper deals with finding two sets of vertices V_p and V'_p that partitions the vertices of the 2D polygon such that they verify Property 1 (the *weak-symmetry* between V_p and V'_p is formally defined in section 3). This step is the most difficult step. Knowing V_p and V'_p the computation of the z -coordinates of the vertices of the 3D polygon is straightforward (the computation method is explained in [2]).

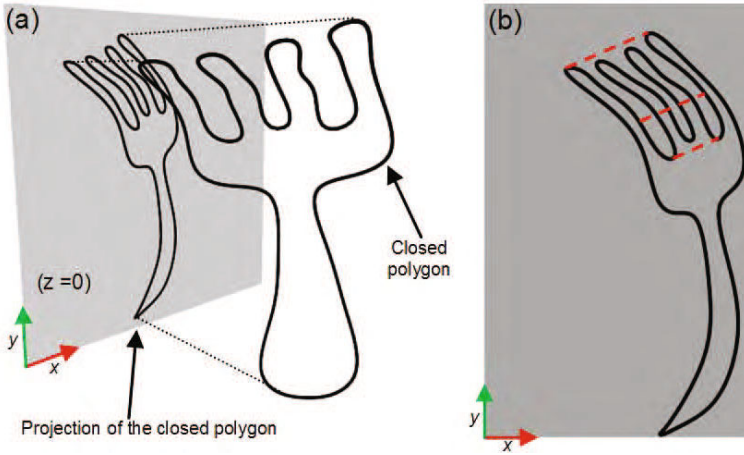


Fig. 1. In (a), the symmetric closed 3D polygon and its orthogonal projection onto the plane $(z=0)$. In (b), the orthogonal projection of the mirror-symmetric 3D polygon; lines joining pairs of symmetric vertices (red dashed lines in the figure) are parallel to each other.

To the best of our knowledge, the symmetry detection from the orthogonal projection of non-planar mirror-symmetric 3D polygons remains an open problem. The closest research work to our approach is the detection of skewed symmetry. Skewed symmetry, as defined by [3], depicts a mirror-symmetric planar curves viewed from some (unknown) viewing direction. Posch [4] has proposed an algorithm for skewed symmetry detection. The algorithm first finds all the segments parallel to the same direction and connecting pairs of symmetric vertices. The skewed symmetry is then detected by checking if the midpoints of these segments are aligned. Shen et al. [5,6] have proposed an algorithm based on an affine-invariant shape representation. They first build a similarity matrix of the vertices of the curves and use this matrix to detect the lines corresponding to the skewed-symmetry axis. Yip [7] has also proposed an approach to detect skewed symmetry axes using Hough transformation. Compared to these previous works, our approach is able to find the weak-symmetry for the projection

of planar and non-planar mirror-symmetric 3D polygons, see Fig. 1. In previous works, the skewed symmetry detection is achieved by finding the symmetry axis. In the case of non-planar mirror-symmetric 3D polygons, such axis does not exist; and thus previous works cannot be used to find the symmetry.

We propose a provable algorithm to the addressed problem of detecting weak-symmetry, To this date no provable algorithm exists to detect the weak-symmetry of a 2D polygon. The presented algorithm comprises two main steps. Firstly, we compute a set of candidate directions which contain every straight line that could make the input polygon weakly-symmetric. Secondly, we take one by one the lines of the candidate-directions set, and we check on the weak-symmetry of the polygon with the help of the sweeping-line strategy [10,9]. In the worst case the whole time complexity of the proposed algorithm is $O(n^3)$.

This article is organized as follows: after the introduction of notation and of the related notions (section 2), section 3 presents the formal definitions of the symmetry problem. Sections 4 and 5 present the steps of the proposed algorithm.

2 Preliminaries

Throughout this article we denote $P = (v_1, \dots, v_n)$ a *simple polygon* of vertices v_i in the counter-clockwise order and of edges the segments $[v_i v_{i+1}]$, where $i = 1, \dots, n$ and $v_{n+1} = v_1$. When P is not a close curve we call P a *polygonal line*. In the following $P(u, v)$ denotes the polygonal line going from the vertex u to the vertex v in the counter-clockwise sense.

Monotonicity: A polygonal line P is said to be *monotone with respect to a straight line ℓ* , if every line parallel to ℓ meets P in at most one point.

A Chain with respect to a direction ℓ : Consider a polygon P , a sub polygonal line $\Gamma \subset P$ is said to be a *chain with respect to a line ℓ* if and only if Γ is monotone with respect to ℓ and satisfies for every monotone polygonal line $\Gamma' \subset P$, if $\Gamma \subset \Gamma'$ then $\Gamma' = \Gamma$. The extremities of the chains of P are called *ℓ -vertices* of P . Fig. 2(a) shows an example of chains and ℓ -vertices.

Type of a vertex: Let ℓ be an oriented line tangent to a vertex $v_i \in P$, v_i is said to be of type **R** if and only if it satisfies one of the following conditions: **(1)** v_{i-1} and v_{i+1} are on the right side of ℓ . **(2)** v_{i-1} is on ℓ and v_{i+1} on the right side of ℓ . **(3)** v_{i+1} is on ℓ and v_{i-1} on the right side of ℓ . **(4)** If v_i has one adjacent vertex then it is on the right side of ℓ . We get the definition of a vertex of type **L** by replacing in the previous definition the term “right” by “left”. Fig. 2(b) shows the different cases and their correspondent type.

3 Weak-Symmetry of a Polygon with Respect to a Line

3.1 Definitions

Defining the neighborhood of a point. Let us denote $]uw[$ the open segment that does not contain its extremities u and w . Let r be a positive real number,

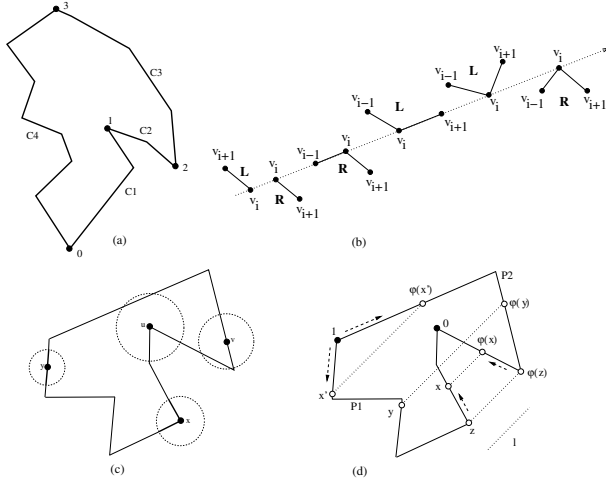


Fig. 2. (a) C1, C2, C3 and C4 are the chains of the polygon with respect to the horizontal. Their extremities are respectively (0, 1), (1, 2), (2, 3) and (3, 0). 0, 1, 2 and 3 are the ℓ -vertices of the polygon (ℓ is the horizontal). (b) Type of a vertex v_i . (c) Examples of neighborhoods of points. Bold lines are neighborhoods of x and y . The discs centered at u and v do not define neighborhoods. (d) A weakly-symmetric polygon. The two polygonal lines of extremities 0 and 1 are weakly-symmetric.

the neighborhood $V(x, r)$ of a point $x \in]v_i v_{i+1}[$ is the intersection set between the closed disc $b(x, r)$ centered at x with P such that $b(x, r) \cap P \subset [v_i v_{i+1}]$. The neighborhood $V(v_i, r)$ of a vertex v_i is the set $b(v_i, r) \cap P$ that verifies $b(v_i, r) \cap P \subset [v_i v_{i-1}] \cup [v_i v_{i+1}]$. The neighborhoods $V(x, r)$ are not defined for every real positive r , however there exists $r_0 > 0$ small enough so that for every $r < r_0$, $V(x, r)$ is well defined, that is, it satisfies the above inclusion constraint. In Fig. 2(c), the intersections of the discs centered at x and y with the polygon define respectively neighborhoods of x and y . However the intersections of the discs centered at u and v with the polygon do not define neighborhoods of u and v .

Definition of the weak-symmetry notion. Two polygonal lines P_1 and P_2 are *weakly-symmetric* with respect to ℓ if and only if there exists a mapping ϕ_ℓ from P_1 to P_2 such that: **(i) Parallel correspondence:** for all $x \in P_1$ the segment $[x\phi_\ell(x)]$ is parallel to ℓ or it is of zero length (i.e $\phi_\ell(x) = x$). **(ii) Bijection:** ϕ_ℓ is bijective. **(iii) Continuity:** ϕ_ℓ and ϕ_ℓ^{-1} are continuous. The continuity of ϕ_ℓ means that for all $x \in P_1$ there exists $\epsilon_0 > 0$ such that for all $\epsilon < \epsilon_0$, there exists $\delta > 0$ so that if $y \in V(x, \delta)$ then $\phi_\ell(y) \in V(\phi_\ell(x), \epsilon)$.

In Fig. 2(d), the polygonal lines P_1 and P_2 (of extremities 0 and 1) are weakly-symmetric with respect to the line ℓ , the weak-symmetry mapping ϕ_ℓ that maps P_1 onto P_2 is defined as follows: We set $\phi_\ell(0) = 0$ and $\phi_\ell(1) = 1$, and we sweep

ℓ over the polygon starting from the vertex 1 and ending at 0. The intersection of ℓ with the polygonal lines $P(1, z)$ and $P(1, \phi_\ell(z))$ and its intersection with $P(z, 0)$ and $P(\phi_\ell(z), 0)$ gives the weakly-symmetric points. For example, the points $\phi_\ell(x')$, $\phi_\ell(y)$ and $\phi_\ell(x)$ are respectively weakly-symmetric to x' , y and x .

A polygon P is *weakly-symmetric with respect to ℓ* if and only if P can be divided in two polygonal lines P_1 and P_2 sharing their extremities and P_1 is weakly-symmetric to P_2 with respect to ℓ . In the rest of the paper, we say *ℓ -weakly-symmetric* instead of “weakly-symmetric with respect to ℓ ”. We say that P is weakly-symmetric if and only if it exists a line ℓ such that P is ℓ -weakly-symmetric.

The polygon of Fig. 2(d) is weakly-symmetric. An example of a polygon which is not weakly-symmetric is illustrated in Fig. 3(a), its non weak-symmetry will be clear after sections 4 and 5.

3.2 Consequences of the Weak-Symmetry Definition

The following Theorem gives two main properties of the weak-symmetry definition. Since there is not enough space, we have chosen not to develop the proofs of Theorems in this article. The reader can find all the proofs and more detailed presentation in the extended version of this paper [8].

Theorem 1. (i) *If a polygon P is weakly-symmetric with respect to ℓ then a vertex of type **R** (respectively **L**) is weakly-symmetric to a vertex of type **R** (respectively **L**).* (ii) *Let $[vv']$ be an edge of P parallel to ℓ , if v is not weakly-symmetric to v' then $[vv']$ is weakly-symmetric to an edge of P aligned with $[vv']$. Otherwise the half-edge $[v \frac{v'+v}{2}]$ is weakly-symmetric to $[v' \frac{v'+v}{2}]$.*

The next sections present the main steps of the algorithm detecting the weak-symmetry of a polygon. Our strategy comprises two main steps. *Step 1 (Finding the candidate directions):* The set of the candidate directions must contain every line, if it exists, that makes the polygon weakly-symmetric. *Step 2 (Verifying the weak-symmetry of a polygon with respect to a candidate direction):* In the set of the candidate directions, we look for a line such that the polygon is weakly-symmetric with respect to this line.

4 The Candidate Directions

4.1 Characterization of the Candidate Directions

Given a polygon P , the goal is to compute a set, as small as possible, containing all the lines such that P is weakly-symmetric with respect to these lines. We call this superset the *candidate directions*. To make this set as small as possible we first present some properties.

Definition 1. We call a concave segment (respectively convex segment) the segment e whose extremities are two concave vertices of P (respectively two convex vertices of P) having the same type with respect to the line passing through e . Also, we call convex-concave segment a segment e whose an extremity is convex, the other is concave and they have the same type with respect to the line passing through e .

An example of convex, concave and convex-concave segments is illustrated in Fig. 3(a). Now, let us define the following sets:

$$\begin{aligned}
 E &= \{e \notin P; e \text{ is a convex - concave segment}\} \\
 E_1 &= E \cup \{e \notin P; e \in CH(P) \text{ or } e \text{ is a concave segment}\} \\
 E_2 &= E \cup \{e \notin P; e \text{ is a convex segment}\}
 \end{aligned}$$

The definition of the candidate directions set, defined in the next definition, is based on the result presented in the following Theorem.

Theorem 2. If a polygon P is ℓ -weakly-symmetric then ℓ is parallel to two edges $e_1 \notin P$ and $e_2 \notin P$ such that $e_1 \in E_1$ and $e_2 \in E_2$.

Definition 4 (the candidate directions). The set of candidate directions is E_1 if the number of the segments in E_1 is smaller than the number of segments in E_2 , otherwise it is E_2 .

We have proven that the number of the candidate directions cannot exceed in the worst case the number $\frac{n(3n-2)}{8}$ (the reader can find the proof in the extended version of this paper [8]).

Example 2. The polygon of Fig. 3(b) has no convex-concave segments. Thus the set E_1 is composed of the segments of extremities taken from the set of vertices $\{2, 4, 6, 8, 10, 12\}$ and the edge of the convex hull $(1, 13)$, the number of segments in E_1 is 16. The set E_2 is a single segment of extremities 1 and 13. Therefore the set of the candidate directions is the set E_2 .

4.2 An Algorithm to Compute the Candidate Directions

We sum up the steps of the algorithm computing the candidate directions set in Algorithms 1 and 2. The time complexity of the algorithm computing the candidate directions is $O(n^2)$ in the worst-case.

```

1 Algorithm 1: Compute-Directions
  input : A polygon  $P$ ,  $m$  convex or concave vertices of  $P$ :  $a_1, \dots, a_m$ .
  output: The set of the directions  $S$ .

2  $S \leftarrow \emptyset$ ,
3  $i \leftarrow 0$ ,
4 while  $i \leq m$  do
5    $j \leftarrow i + 1$ 
6   while  $j \leq m$  do
7      $e \leftarrow [a_i a_j]$ 
8     if  $a_i$  and  $a_j$  have the same type with respect to the direction  $e$  and  $e$ 
       is not parallel to an edge of  $S$  then
9       | add  $e$  to  $S$ 
10      end
11       $j \leftarrow j + 1$ 
12    end
13     $i \leftarrow i + 1$ 
14 end

```

Algorithm 1. Computing the sets E_1 and E_2

```

1 Algorithm 2: Candidate-Directions
  input : A polygon  $P$ .
  output: list of the candidate directions

2 Compute  $CH(P)$  the convex hull of  $P$ . Let us  $a_1, \dots, a_{n_1}$  be the concave
  vertices of  $P$  and  $b_1, \dots, b_{n_2}$  are the convex vertices of  $P$ .
3  $E_1 \leftarrow$  Algorithm 1: Compute-Directions ( $P, a_1, \dots, a_{n_1}$ )
4  $E_2 \leftarrow$  Algorithm 1: Compute-Directions ( $P, b_1, \dots, b_{n_2}$ )
5 Compute the convex-concave segments of  $E$ 
6 Add the edges of  $CH(P) - P$  and the segments of  $E$  to  $E_1$  and to  $E_2$ 
7 if  $|E_1| < |E_2|$  then
8   | return  $E_1$ 
9 end
10 else
11 | return  $E_2$ 
12 end

```

Algorithm 2. Computing the candidate directions

5 Filtering the Set of the Candidate Directions

Our goal in this section is to select the right direction from the candidate set, that is for each line of this set we verify that the polygon P is either ℓ -weakly-symmetric or not. To do this, our algorithm sweeps ℓ over P . Without loss of generality, we suppose that the direction ℓ is horizontal, the algorithm comprises the following two main steps.

Step 1: Initialization (compute two ℓ -weakly-symmetric vertices). To start the process of sweeping the horizontal over the polygon, we need two weakly-symmetric vertices. These vertices are the lowest leftmost and rightmost vertices. This claim is presented in Theorem 3, The proof is presented in the extended version [8]. Consider the lowest vertices $u_1 < u_2 < \dots < u_r$ of P sorted according to their increasing abscissa. An example of such vertices is $0 < 4 < 10$ of the polygon illustrated by Fig. 3(c).

Theorem 3 (First weakly-symmetric vertices). *If the polygon P is ℓ -weakly-symmetric then the weakly-symmetric of u_1 is u_r .*

Step 2: The sweeping-line process. Recall that $P(u, v) \subset P$ is the polygonal line extracted from P , its extremities are the vertices u and v , the other vertices of the polygonal line are obtained by scanning P in the clockwise orientation from u and counter-clockwise from v . We divide P on two polygonal lines, $P_1 = P(u_1, u_r)$ and $P_2 = P(u_r, u_1)$ and we check on the ℓ -weak-symmetry of P_1 and P_2 . Beginning from u_1 and u_r , we sweep the line over P_1 and then over P_2 . The sweep stops at discrete “events” that is the line ℓ hits ℓ -vertices or edges parallel to ℓ . We check on that the touched ℓ -vertices have the same type or the line ℓ contains weakly-symmetric edges parallel to ℓ (see Theorem 1). If once this property is not satisfied we reject ℓ , otherwise we return that P is ℓ -weakly-symmetric.

The key step of the sweeping-line process is the verification of the weak-symmetry of a polygonal line, it is summed up in Algorithm 3. The input of Algorithm 3 is a polygonal line $P(v_i, v_j)$, where v_i and v_j are on the horizontal. The weak-sweeping process will take the clockwise order from v_i and the counter-clockwise order from v_j . $next(j)$ is the label of the next ℓ -vertex in the counter-clockwise order, $previous(i)$ is the label of the next ℓ -vertex in the clockwise sense. The following algorithm returns true if the polygonal line $P(v_i, v_j)$ is weakly-symmetric.

Example 3. (Steps of the algorithm 3 through the example of Fig. 3(c)).

Applying this algorithm to the polygonal line $P(10, 0) = (0, 1, 2, \dots, 8, 9, 10)$ of the polygon of Fig. 3(c), we get these iterations. Since 0 and 10 are weakly-symmetric, (Theorem 3), the function $previous(10)$ returns 7, and $next(0)$ returns the vertex 2. Since 2 and 7 are weakly-symmetric, the sweep continue: $previous(7)$ looks for the next ℓ -vertex in the clockwise sense it is 4, $next(2)$ returns also 4. Since $v_i = v_j = 4$ the loop stops and returns that the polygonal line $P(10, 0)$ is weakly-symmetric. Let us turn to the polygonal line $P(0, 10) = (10, 11, \dots, 29, 0)$. The iterations of the loop while are: firstly, $previous(0)$ returns 26 and $next(10)$ returns 12, since they are weakly-symmetric, the next iterations check the weak-symmetry respectively between 25 and 13, 24 and 14. When it reaches 22 and 16, the algorithm stops and returns false because these two vertices are not weakly-symmetric.

Example 4. (Steps of the filtering algorithm through the example of Fig. 3(c)). First of all, we transform the polygon P such that the direction ℓ becomes horizontal. The lowest vertices u_i allows us to initialize the sweep process. In Fig. 3(c), the ordered vertices u_i are $0 < 4 < 10$, thus the vertices 0 and 10 are weakly-symmetric. we apply Algorithm 3: weakly-symmetric-polygonal line to $P(10, 0)$, it returns true, therefore we call again Algorithm3: weakly-symmetric-polygonal line for the polygonal line $P(0, 10)$, it returns false since it fails at the vertices 22 and 16. Thus P is not weakly-symmetric with respect to the input direction.

```

1 Algorithm 3: weakly-symmetric-polygonal-line
   input : The polygonal line  $P(v_i, v_j)$ 
   output: return true if  $P(v_i, v_j)$  is weakly-symmetric otherwise it returns
           false
2  $i \leftarrow previous(i). j \leftarrow next(j)$ 
3 weakly-symmetric  $\leftarrow$  true
4 while ( $v_i \neq v_j$  and  $([v_i v_j] \notin P$  and  $[v_i v_j]$  is horizontal) and
   weakly-symmetric) do
5   if  $v_i$  is weakly-symmetric to  $v_j$  then
6     |  $i \leftarrow previous(i). j \leftarrow next(j)$ 
7   end
8   else
9     | weakly-symmetric  $\leftarrow$  false
10  end
11  Return weakly-symmetric
12 end

```

Algorithm 3. Verifying the weak-symmetry of the polygonal-line

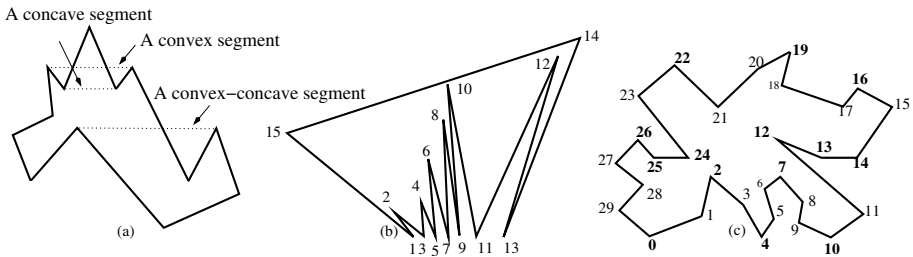


Fig. 3. (a) An example of convex, concave and convex-concave segments. (b) A polygon with one convex segment: (1, 13) and 15 concave segments. This polygon is not weakly-symmetric. (c) A non weakly-symmetric polygon with respect to the horizontal. The bold vertices are the ℓ -vertices, The vertex 22 is not weakly-symmetric to the vertex 16.

6 Conclusion

Detecting the weak-symmetry of a planar hand sketched polygon is a key step to reconstruct mirror-symmetric non-planar 3D-polygons. We have formalized the notion of weak-symmetry and have proposed a provable solution. No such algorithm is known till date. For an input simple polygon, the presented algorithm computes first a small set that contains all the direction that could make the polygon weakly-symmetric. Secondly we iteratively look for, in this set, the direction of weak-symmetry. The whole time complexity of the algorithm is $O(n^3)$ in the worst case. The ongoing work will be the extension of this approach to check on the weak-symmetry of a collection of non simple polygons.

References

1. Olsen, L., Samavati, F.F., Costa Sousa, M., Jorge, J.-A.: Sketch-based modeling: A survey. *Computers & Graphics* 33(1), 85–103 (2009)
2. Cordier, F., Seo, H., Park, J., Noh, J.: Sketching of Mirror-symmetric Shapes. *IEEE Transactions on Visualization and Computer Graphics* (2011)
3. Kanade, T.: Recovery of the Three-Dimensional Shape of an Object from a Single View. *Artificial Intelligence* 17, 409–460 (1981)
4. Posch, S.: Detecting skewed symmetries. In: *International Conference on Pattern Recognition*, The Hague, pp. 602–606 (1992)
5. Shen, D., Horace, H.-S Ip., Teoh, E.K.: Robust detection of skewed symmetries by combining local and semi-local affine invariants. *Pattern Recognition* 34(7), 1417–1428 (2001)
6. Shen, D., Horace, H.-S Ip., Teoh, E.K.: Robust Detection of Skewed Symmetries. In: *International Conference on Pattern Recognition*, pp. 7022–7025 (2000)
7. Yip Raymond, K.-K.: A Hough transform technique for the detection of reflectional symmetry and skew-symmetry. *Pattern Recognition Letters* 21(2), 117–130 (2000)
8. Melkemi, M., Cordier, F., Sapidis, N.: A provable algorithm to detect the symmetry of a simple polygon. *Technical Report LMIA*, pp. 1–30 (2010)
9. Okabe, A., Boots, B., Sugihara, K.: *Spatial tessellations: Concepts and Applications of Voronoi Diagrams*. John Wiley and Sons, Chichester (1992)
10. O'Rourke, J.: *Computational geometry in C*. Cambridge University Press, Cambridge (1998)

Spatially Variant Dimensionality Reduction for the Visualization of Multi/Hyperspectral Images

Steven Le Moan^{1,2}, Alamin Mansouri¹, Yvon Voisin¹, and Jon Y. Hardeberg²

¹ Le2i, Université de Bourgogne, Auxerre, France

² Colorlab, Høgskolen i Gjøvik, Norway

Abstract. In this paper, we introduce a new approach for color visualization of multi/hyperspectral images. Unlike traditional methods, we propose to operate a local analysis instead of considering that all the pixels are part of the same population. It takes a segmentation map as an input and then achieves a dimensionality reduction adaptively inside each class of pixels. Moreover, in order to avoid unappealing discontinuities between regions, we propose to make use of a set of distance transform maps to weigh the mapping applied to each pixel with regard to its relative location with classes' centroids. Results on two hyperspectral datasets illustrate the efficiency of the proposed method.

1 Introduction

Spectral imagery consists of acquiring a scene at more than three different ranges of wavelengths, usually dozens. Since spectral display devices are yet rare, most of today's popular display hardware is based on the tri-stimulus paradigm [1]. Thus, in order to visualize spectral images, a dimensionality reduction step is required so that only three channels (Red, Green and Blue for example) can contain most of the visual information while easing interpretation by preserving natural colors and contrasts [2]. At this aim, many dimensionality reduction techniques have been applied to the task of visualizing spectral datasets, they are roughly divided into two categories: either they operate a transformation or a selection of spectral channels. Even though the latter family is a subset of the former one, they are based on two very different philosophies. Indeed, band selection aim at preserving the physical meaning of spectral channels by keeping them intact during the N -to-3 projection, whereas band transformation allows any combination of channels (even nonlinear) as a means to fuse information along the spectrum. Therefore, the choice between these two approaches is of course application-driven. Band transformation methods are, for instance, based on the use of Principal Components Analysis (PCA) [3,4], Color Matching Functions (CMF) [5,2] or Independent Components Analysis [6]. Band selection strategies involve the use of similarity criteria such as correlation [7], Mutual Information [8] or Orthogonal Subspace Projection [9]. All these methods are based on the assumption that all the pixels are part of the same population, i.e. they perform a global mapping. Scheunders [10] proposed to spatially divide

the image into blocks in order to achieve local mappings by means of PCA and Neural Network-based techniques. Discontinuities between blocks are dealt with by adapting the mappings at a pixel level. Not only do we propose to extend Scheunders' approach from a greyscale to a color framework, we enhance it in two ways: by using a classification map so as to choose which visual features deserve a local contrast enhancement, and by introducing a weighing function allowing to balance not only the influence of global versus local mapping, but also the respective influences of the individual classes. We will first introduce the different steps of the proposed approach: classification map, distance transforms and weighing of dimensionality reduction functions. Results will then be presented and discussed before conclusion.

2 Spatially Variant Dimensionality Reduction

In this section, we give details on the different elements involved in the procedure.

2.1 Segmentation Map

The first step of the proposed technique is to obtain a spatial segmentation of the image. This can be achieved either manually or automatically, by means of classifiers such as the K-Means, or Support Vector Machines. The choice of such a method is considered outside the scope of this paper as long as it is application-dependent and that the following processings apply anyway.

Let then $\mathbf{Seg}_K(I)$ be a segmentation map of image I containing K classes. While traditional methods consider each spectral channel as a whole, the core idea of the spatially variant dimensionality reduction is to analyze sets of pixels independently. For instance, if one desires to enhance the contrast between a couple of specific objects, one must consider the corresponding set of pixels separately from the others, in order to obtain a more dedicated analysis. Therefore, the final segmentation map must be computed not in a way that similar pixels are clustered together, but so that each class contains objects that need to be "separated".

2.2 Distance Transform

In order to locally adapt the dimensionality reduction so that no discontinuities occur between regions, we need to know, for each location in the image, the distance to the closest centroids of each class. The distance transform is a way to efficiently achieve such measurements. It applies on binary images and consists of computing for each pixel with value 0 (black), its distance to the closest one with value 1 (white). Therefore, we need a set of binary images containing, in white, all the centroids of the different connected components from class C_i and all the other pixels in black. We obtain such distance transform maps as the ones depicted in Figure 1c and 1f. Eventually, for a pixel $p_{(x,y)}^{c_i}$ at spatial coordinates (x, y) , belonging to class c_i , we obtain the set of its respective distances to the other classes centroids $\mathbf{d}(x, y) = [d_1(x, y), \dots, d_K(x, y)]^T$, including the distance to the closest centroid of its own class.

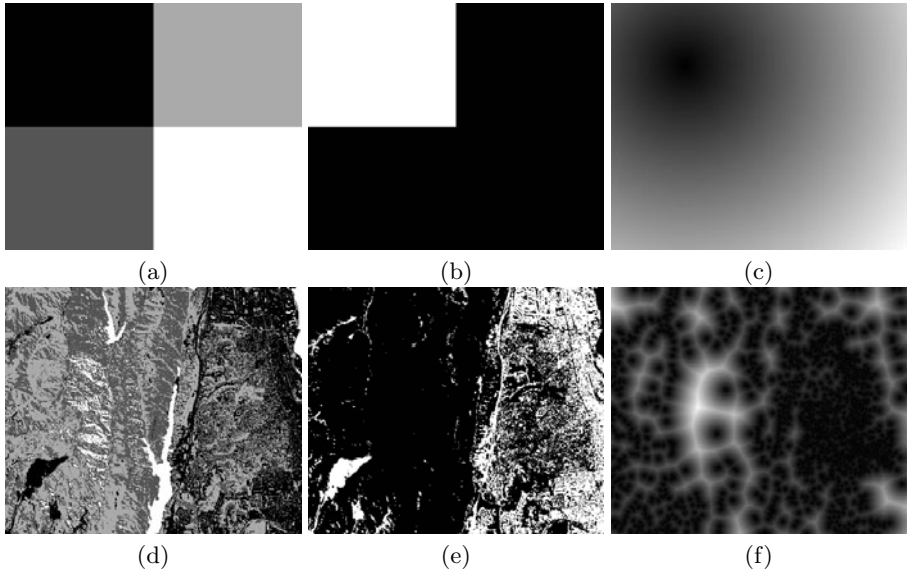


Fig. 1. Illustration of the distance transform applied on two possible segmentation maps for the "Jasper Ridge" dataset (see results section for full description). First column: segmentation maps (4 classes), Second column : class 1 isolated in white, Third column: the corresponding distance transforms. The first segmentation has been achieved manually, whereas the second one is the result of the K-means classifier.

2.3 Weighing of Dimensionality Reduction Functions

Dimensionality Reduction (DR) is then performed in each class independently from the others so that we obtain as many sets of DR functions as there are classes. Moreover, a global mapping is also performed so as to be able to further balance between global and local mapping. What we refer to as a DR function is nothing more than a vector of coefficients used for fusing the spectral channels in order to obtain one of the three (Red, Green or Blue) primary bands. For instance, the Color Matching Functions (CMF) are such vectors.

Each pixel is then being affected with a set of weighted DR functions $DR^{Red}(x, y)$, $DR^{Green}(x, y)$ and $DR^{Blue}(x, y)$ such that:

$$DR^{Red}(x, y) = \omega_0 \times \mathbf{DR}_0^{Red} + (1 - \omega_0) \times \frac{\sum_{k \in \{1..K\}} \omega_k \times d_k(x, y) \times \mathbf{DR}_k^{Red}}{\sum_{k \in \{1..K\}} d_k(x, y)}$$

with ω_0 being the parameter allowing to balance between global and local mappings and $\omega = [\omega_1, \dots, \omega_K]^T$, the vector of coefficients depicting the respective influences of the classes (its sum must be equal to one). The latter can be set manually or automatically, so that, for example, largest classes are given more weight. \mathbf{DR}_0^{Red} is the global DR function and $\mathbf{DR}_k^{Red}, \forall i \in [1..K]$ are the local ones. Similar definitions apply of course for $DR^{Green}(x, y)$ and $DR^{Blue}(x, y)$.

3 Experiments and Results

3.1 Data Sets

For our experiments, we have used two hyperspectral datasets:

- "Jasper Ridge" is a well-known 220 bands image from the AVIRIS sensor [11]. We have used only a portion of the original dataset for the sake of clarity.
- "Norway" is a 160 bands remote sensing image, representing a urban area in the neighborhood of Oslo (Norway). It was acquired with the HySpex VNIR-1600 sensor, developed by the Norsk Elektro Optikk company in Oslo. The sensor ranges from the early visible (400nm) to the near infrared (1000nm) with a spectral resolution of 3.7 nm [12].

As a pre-processing step, bands with average reflectance value below 2% and those with low correlation (below 0.8) with their neighboring bands have been removed, as suggested in [13].

3.2 Dimensionality Reduction Techniques

We have selected two dimensionality reduction techniques to illustrate the proposed approach.

- PCA_{hsv} is the traditional Principal Components Analysis of which components are mapped to the HSV color space ($PC1 \rightarrow V; PC2 \rightarrow S; PC3 \rightarrow H$).
- LP is a state-of-the-art band selection approached which has been proposed by Du *et al.* [9] and consists of progressively selecting bands by maximizing their respective orthogonality.

3.3 Evaluation

In order to evaluate the improvements by the proposed approach, we have, based on a K-means classification, selected 3 objects (or classes) of interest (Obj_1, Obj_2 and Obj_3) in each image and used the color difference metric CIE76 ΔE_{ab^*} as a means to measure how contrasted they are. Obviously, the more they are contrasted, the more visual information we have. This metric has been applied on the objects' centroids (in the color space CIELAB). The objects of interest are depicted on Figure 2 for both datasets. For the spatially-variant dimensionality reduction, we have then used a segmentation map so that we obtain the three classes : $C_1 = \{Obj_1 \cup Obj_2\}$, $C_2 = Obj_3$ and $C_3 = \{\text{rest of the pixels}\}$, as shown in Figures 2b and 2d.

The object-separability metric will be referred to as Inter-Object Perceptual Separability (IOPS)

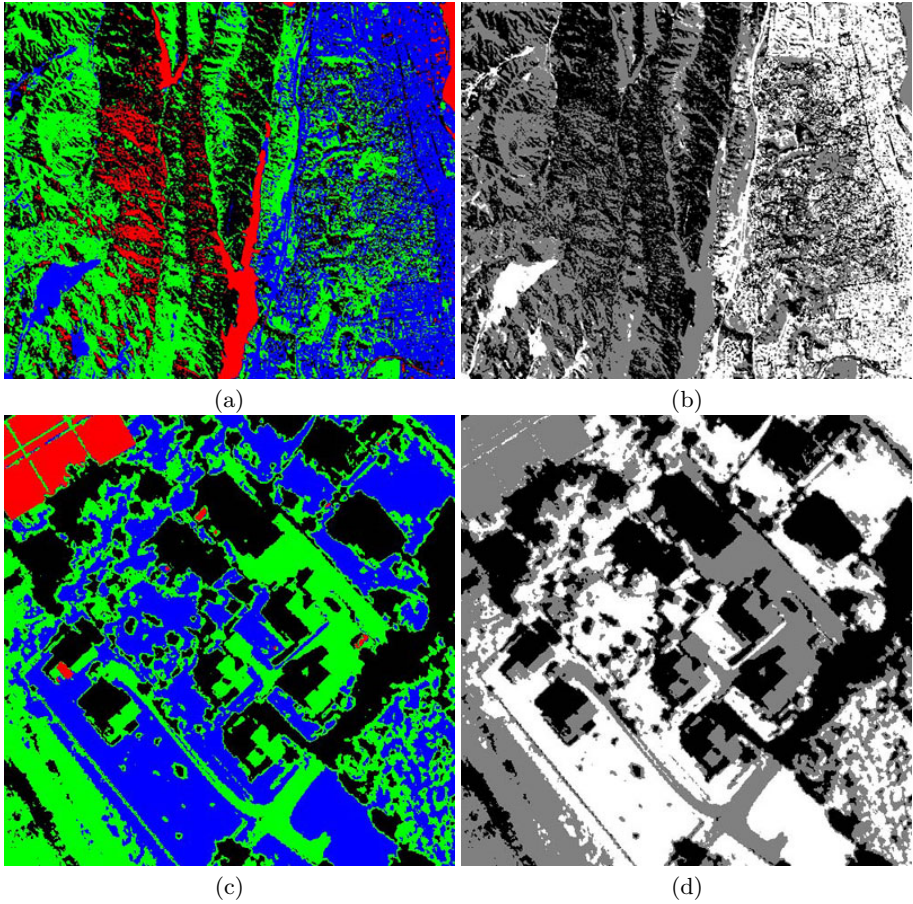


Fig. 2. Selected classes of interest (Obj_1 in red, Obj_2 in green and Obj_3 in blue)

3.4 Results

Figures 3 and 4 depict respectively the resulting color composites by the global mappings ($\omega_0 = 1$) and the local mapping ($\omega_0 = 0$) without smoothing. Figure 5 depicts the obtained color composites while Tables 1 and 2 give the results in terms of IOPS for the following configurations:

- Config 1: $\omega_0 = 0$ and $\omega = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$
- Config 2: $\omega_0 = 0$ and $\omega = [0.8, 0.1, 0.1]$
- Config 3: $\omega_0 = 0$ and $\omega = [0.1, 0.8, 0.1]$
- Config 4: $\omega_0 = 0.5$ and $\omega = [\frac{1}{3}, \frac{1}{3}, \frac{1}{3}]$

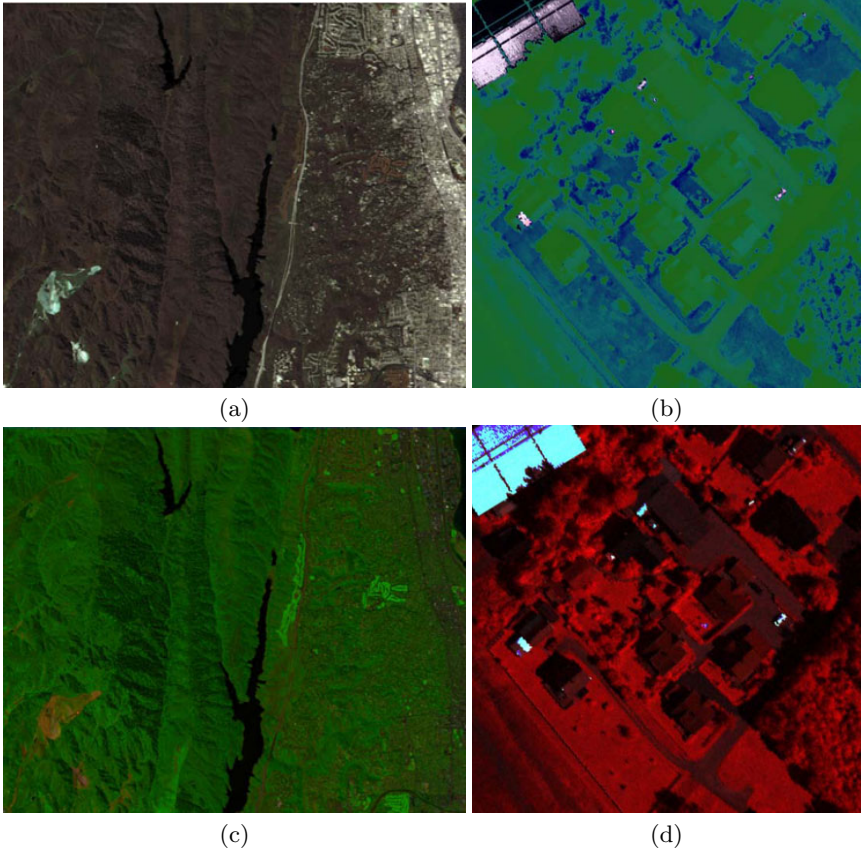


Fig. 3. Results obtained for $\omega_0 = 1$ (global only). First row: $PCA_{h,sv}$, second row: LP-based band selection.

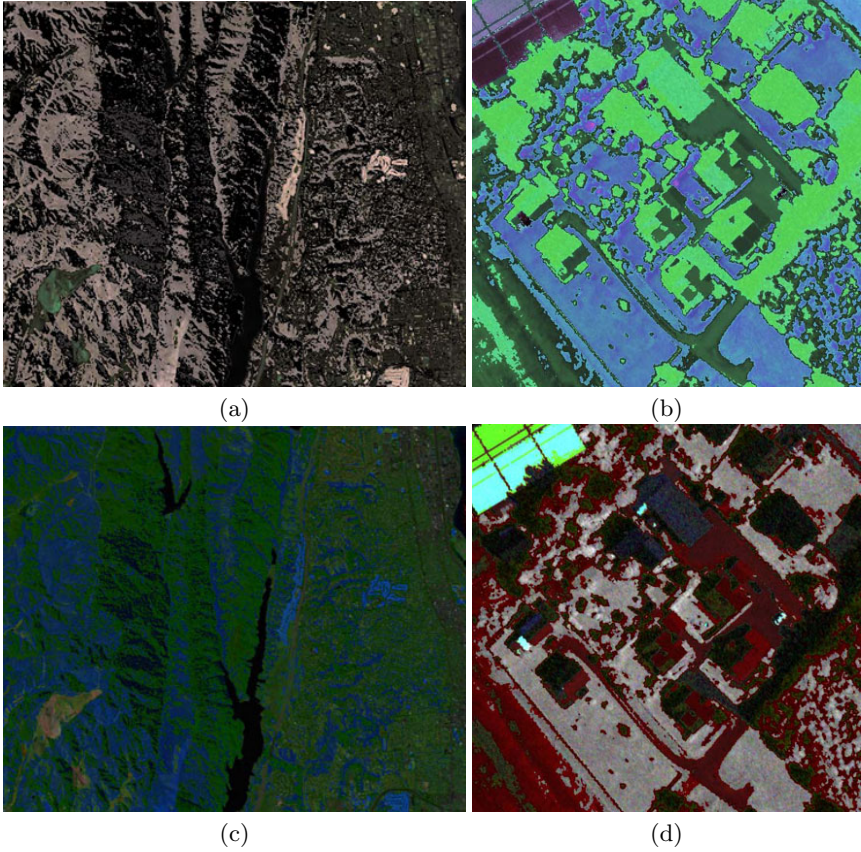


Fig. 4. Results obtained for $\omega_0 = 0$ and without weighing of the DR functions. First row: PCA_{hsv} , second row: LP-based band selection.

Table 1. Inter-Object Perceptual Distance results for the "Jasper Ridge" image and for all the configurations considered

		PCA_{hsv}	LP
Global	Obj_1 vs. Obj_2	19.3	21.4
	Obj_1 vs. Obj_3	31.0	12.2
Config 1	Obj_1 vs. Obj_2	35.7	34.3
	Obj_1 vs. Obj_3	21.2	11.3
Config 2	Obj_1 vs. Obj_2	44.4	39.0
	Obj_1 vs. Obj_3	26.1	11.0
Config 3	Obj_1 vs. Obj_2	34.0	31.5
	Obj_1 vs. Obj_3	21.3	10.3
Config 4	Obj_1 vs. Obj_2	27.3	30.9
	Obj_1 vs. Obj_3	26.2	11.2

Table 2. Inter-Object Perceptual Distance results for the "Norway" image and for all the configurations considered

		PCA_{hsv}	LP
Global	Obj_1 vs. Obj_2	44.8	65.3
	Obj_1 vs. Obj_3	24.2	37.0
Config 1	Obj_1 vs. Obj_2	56.0	68.0
	Obj_1 vs. Obj_3	21.3	27.7
Config 2	Obj_1 vs. Obj_2	70.3	74.5
	Obj_1 vs. Obj_3	22.1	27.6
Config 3	Obj_1 vs. Obj_2	60.1	71.2
	Obj_1 vs. Obj_3	24.0	28.6
Config 4	Obj_1 vs. Obj_2	48.8	73.4
	Obj_1 vs. Obj_3	22.7	31.2

4 Comments on the Results

Based on the results presented in the previous section, we make the following remarks:

- The absence of weighing of the DR functions results in sharp discontinuities, as one can notice on Figures 4. Such artifacts are quite unappealing and thus do not allow for an efficient interpretation, hence the usefulness of the smoothing achieved by the weighing of the DR functions.
- Overall, one can observe significant improvements from global to local techniques in the separation of Objects 1 and 2. On the other hand, separation between Objects 1 and 3 is better handled by the global approach. This comes from the fact that those three objects are given the same mapping in the global configuration, unlike in the local one.
- The second configuration always gives the best separation between Objects 1 and 2. This is due to the fact that this configuration gives more weight to the DR achieved in the class formed by these objects.

- The global DR allows for an overall better separation but local contrasts are not optimized, since better results are obtained from the first configuration, where all the classes are considered of equal influence. As a follow to that comment, the fourth configuration, which uses a 50-50 combination of global and local mappings gives the best compromise between both separations.

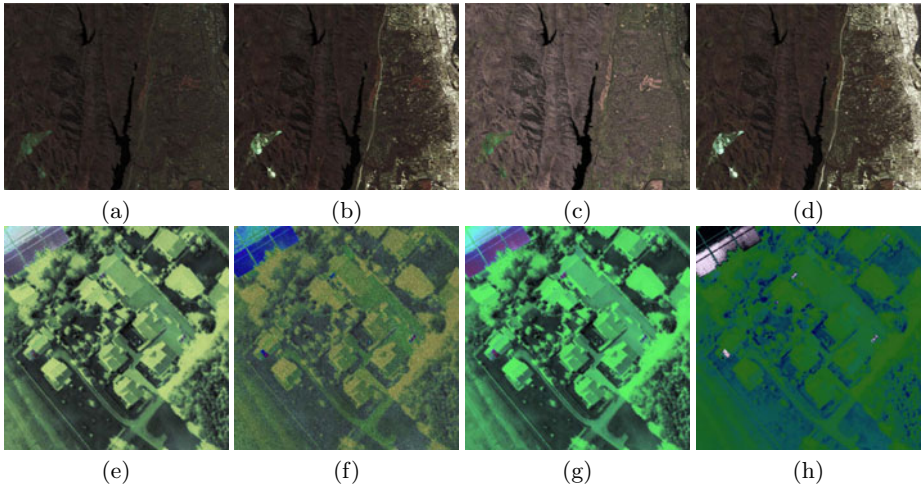


Fig. 5. Results obtained with the PCA-based dimensionality reduction, for all the configuration - First column: Config 1, second column: Config 2 , third column: Config 3,, fourth column: Config 4

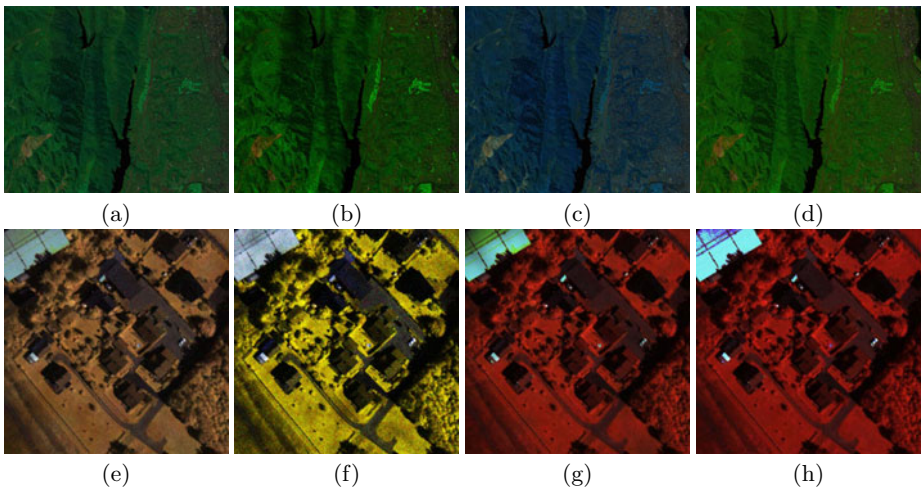


Fig. 6. Results obtained with the LP-based dimensionality reduction, for all the configuration - First column: Config 1, second column: Config 2 , third column: Config 3,, fourth column: Config 4

5 Conclusions

An adaptive feature extraction algorithm has been presented, which takes into account dissimilarities between pixels by first clustering them and then conducting dimensionality reduction separately in each cluster. Preliminary results show an increasing amount of informative as well as perceptual content. The technique being obviously very sensitive to the pixel clustering conducted prior to dimensionality reduction, this step will be further investigated along with the influence of other feature extraction techniques as well as other distance metrics in order to draw a more complete evaluation of the spatially-variant dimensionality reduction.

Acknowledgements

The regional council of Burgundy supported this work.

References

1. Grassmann, H.: On the theory of compound colors. *Phil. Mag.* 7, 254–264 (1854)
2. Jacobson, N., Gupta, M.: Design goals and solutions for display of hyperspectral images. *IEEE Trans. on Geoscience and Remote Sensing* 43, 2684–2692 (2005)
3. Jia, X., Richards, J.: Segmented principal components transformation for efficient hyperspectral remote-sensing image display and classification. *IEEE Trans. on Geoscience and Remote Sensing* 37, 538–542 (1999)
4. Tyo, J., Konsolakis, A., Diersen, D., Olsen, R.: Principal-components-based display strategy for spectral imagery. *IEEE Trans. on Geoscience and Remote Sensing* 41, 708–718 (2003)
5. Poldera, G., van der Heijden, G.: Visualization of spectral images. In: *Proc. SPIE*, vol. 4553, p. 133 (2001)
6. Hyvärinen, A., Oja, E.: Independent component analysis: algorithms and applications. *Neural Networks* 13, 411–430 (2000)
7. Chang, C., Du, Q., Sun, T., Althouse, M.: A joint band prioritization and band-decorrelation approach to band selection for hyperspectral image classification. *IEEE Trans. on Geoscience and Remote Sensing* 37, 2631–2641 (1999)
8. Guo, B., Damper, R., Gunn, S., Nelson, J.: A fast separability-based feature-selection method for high-dimensional remotely sensed image classification. *Pattern Recognition* 41, 1670–1679 (2008)
9. Du, Q., Yang, H.: Similarity-based unsupervised band selection for hyperspectral image analysis. *IEEE Geoscience and Remote Sensing Letters* 5, 564–568 (2008)
10. Scheunders, P.: Multispectral image fusion using local mapping techniques. In: *International Conference on Pattern Recognition*, vol. 15, pp. 311–314 (2000)
11. <http://aviris.jpl.nasa.gov/html/aviris.freedata.html> (last check: November 11, 2010)
12. <http://www.neo.no/hyspex/> (last check: November 11, 2010)
13. Cai, S., Du, Q., Moorhead, R.: Hyperspectral imagery visualization using double layers. *IEEE Trans. on Geoscience and Remote Sensing* 45, 3028–3036 (2007)

Maneuvering Head Motion Tracking by Coarse-to-Fine Particle Filter

Yun-Qian Miao¹, Paul Fieguth², and Mohamed S. Kamel¹

¹ Department of Electrical and Computer Engineering, University of Waterloo,
Waterloo, Ontario, Canada, N2L 3G1
{mikem,mkamel}@pami.uwaterloo.ca

² Department of System Design Engineering, University of Waterloo, Waterloo,
Ontario, Canada, N2L 3G1
pfieguth@uwaterloo.ca

Abstract. Tracking a very actively maneuvering object is challenging due to the lack of state transition dynamics to describe the system's evolution. In this paper, a coarse-to-fine particle filter algorithm is proposed for such tracking, whereby one loop of the traditional particle filtering approach is divided into two stages. In the coarse stage, the particles adopt a uniform distribution which is parameterized by the limited motion range within each time step. In the following fine stage, the particles are resampled using the results of the coarse stage as the proposal distribution, which incorporates the most present observation. The weighting scheme is implemented using a partitioned color cue that implicitly embeds geometric information to enhance robustness. The system is tested by a publicly available dataset for tracking an intentionally erratic moving human head. The results demonstrate that the proposed system is capable of handling random motion dynamics with a relatively small number of particles.

Keywords: motion tracking, particle filter, color cue, coarse-to-fine.

1 Introduction

With the availability of high-power computers, inexpensive video cameras and the increasing needs, object tracking is one active research area and is required in many applications such as automated surveillance, traffic monitoring, human-computer interfaces, and other motion-based recognitions [15].

In object tracking applications, the most popular methods for estimating target positions incorporate variations of the Kalman filter [14][12]. Under the assumption of a linear system with Gaussian noise, the Kalman Filters (KFs) achieve the optimal estimation in terms of the minimal covariance.

However, in cases where the target is actively maneuvering, the system evolution model is far from linearity and even shows multi-modality, and the performance of the Kalman Filter degrades unsurprisingly. A typical example is tracking people, whose erratic movements are poorly matched to any model of more than second order [1].

The particle filter (PF) is the result of applying the Monte Carlo method in recursive Bayesian filtering [13], which is used to estimate the posterior Probability Density Function (PDF) of the state variable based on Bayes Theorem. In a particle filter, the PDF is sampled and represented by a collection of particles with weights proportional to their likelihood. This representation of the distribution makes no assumption about the distribution shape, and is capable of handling non-linear systems with non-Gaussian and multimodal distributions.

In this paper, a coarse-to-fine approach is proposed to enhance the particle filter by dividing one loop of particle filter into two stages. For tracking highly varied dynamics system, human motion, we cannot model the system dynamics in a simple form, such as linear or Gaussian. Instead, we may just use the prior knowledge about the dynamic range limits on the system state transition function. Therefore, in the coarse stage, the uniform distribution with range limits is used as the proposed distribution. In the following fine stage, only the highly probability area that are discovered by the coarse stage is explored further with a Gaussian model. In this tracking system, we adopt partitioned color cue that embeds geometric information implicitly as particle's weighting scheme.

The proposed algorithm is tested with a challenging video sequence for the task of tracking an intentional randomly moving human head. The experimental results from both observations and quantitative results suggest that the coarse-to-fine PF is capable of handling the tracking high varied motions with relative small number of particles.

The outline of this paper is as follows. Section 2 reviews backgrounds on the tracking problem itself and the basic form of particle filters. Section 3 introduces a coarse-to-fine approach and also explains the weighting mechanism that is taken in our tracking system. Experiments on a real video sequence are presented in section 4. The paper is summarized with conclusions and future directions in section 5.

2 Background

2.1 What to Track

The first question of the tracking problem is the choice of what to track. This is close to the features selected for representing the interested object. In a visual tracking system, objects are usually represented by their shapes and appearances.

Object representations are usually chosen according to different application scenarios. For tracking very small objects in an image, point representation is usually appropriate. For example, Veenman et al. [13] use the point representation to track the seeds in a moving dish sequence. Similarly, Shafique and Shah [10] use the point representation to track distant birds. For objects whose shapes can be approximated by rectangles or ellipses, primitive geometric shape representations are more suitable. Shen et al. [11] use an elliptical shape representation integrated with color histogram computed from the elliptical region for modeling the appearance of face. For tracking objects with complex shapes,

for example, humans, a contour or a silhouette-based representation is common [7]. Practically, many tracking algorithms use a combination of multiple cues to achieve more reliable results [11].

2.2 State Space and the Dynamic Model

The object of interest can be formulated by the state sequence $\{z_t, t = 1, 2, \dots\}$, which evolves along the time domain. For a discrete time step, when using an ellipse or rectangle model in the image space domain, the object is parameterized by:

$$z = \{x, y, a, b\} \quad (1)$$

where x and y denote the centroid of the ellipse or rectangle, a and b denote the length of the half axes.

Another key problem is the representation of the system dynamics. The general form of first-order Markov system evolving model is:

$$z_t = f(z_{t-1}, w_{t-1}) \quad (2)$$

where $f()$ is the state transition function, z_t is the system state which we are interested in and w_t is the process noise sequence. Then, the objective of tracking is to recursively estimate z_t by giving a sequence of measurements m_t , where the measuring model is:

$$m_t = h(z_t, v_t) \quad (3)$$

where $h()$ is the measurement model and v_t is a sequence of observation noise.

2.3 The Particle Filter

Following the above system state expressions, by simplifying $f()$ and $h()$ to be linear and w, v to be Gaussian, the Kalman Filters (KFs) achieve optimal estimations. In this case, the above equations (2) and (3) are simplified to:

$$z_t = Az_{t-1} + w_{t-1} \quad (4)$$

$$m_t = Cz_t + v_t \quad (5)$$

For realistic systems that do not ideally follow these assumptions, some variants of KF were developed in the literature, such as Extended Kalman Filter (EKF) [1], unscented Kalman filter (UKF) [5], ensemble Kalman Filter (EnKF) [4].

If a system is far from the linear and Gaussian model, instead of analytically solving the tracking problem, the particle filters take a different approach that uses the sampling method for approximating any general form of distribution. From the Bayesian perspective, if we can recursively calculate the posterior probability density function (PDF), $p(z_t|m_{1:t})$, we can easily conduct our estimation based on it, such as expectation or maximum a posterior probability (MAP).

3 Proposed Approach

For tracking actively maneuvering motions, such as erratic movement of the human head, the system state's transition needs to be considered carefully. Due to the fast varying in the moving velocity, acceleration, and dramatically changing directions, using the linear model to represent the state transition is not suitable. We present an amended approach that divided one loop of particle filter into a coarse stage and a fine stage.

3.1 Overview of the Coarse-to-Fine Particle Filter

In [12], this problem is noticed and the authors break the iteration by inserting a motion estimation step in the main particle filter loop. However, their proposed method is still based on linear and Gaussian models for both the coarse and fine steps, just with adjustments of different covariance for a sequence of three stages in one loop.

Instead, based on re-examining the system state transition function $z_t = f(z_{t-1}, w_{t-1})$, we cannot conclude a linear function for f if the motion of an object is varying randomly. But, this does not mean we can't describe the system's evolving state. We still can infer some knowledge about the dynamics of system. Here, for tracking a human who is intentionally moving randomly, we can apply the range limitation about the f , that is:

$$\|z_t - z_{t-1}\| \leq R \quad (6)$$

Further, this infers the system dynamics as:

$$z_t \in z_{t-1} \pm R \quad (7)$$

where R is the pre-setting that explains the limitation of moving range for the time step duration.

As a result, in the coarse step we use a uniform distribution that ranges in the scope of $z_{t-1} \pm R$. Following, the first round of weighting and resampling are executed. Then, in the fine step we can focus on the high likelihood area implicitly. In the fine step, we adopt Gaussian model and use small variance to limit searching scope to expect a better performance.

In fact, this coarse-to-fine approach is intended to use the optimal proposal distribution $p(z_t|z_{t-1}, m_t)$ instead of $p(z_t|z_{t-1})$, which incorporates the newest measurement into the proposal distribution.

At the end of the fine step for each frame of image, the estimated state is produced by choosing the hypothesis with the maximum weight following the MAP principle.

The details of the proposed coarse-to-fine particle filter (CFPF) are described in Algorithm 1.

Algorithm 1. Motion tracking of Coarse-to-Fine particle filter

Require: state of $t - 1$

1. Coarse stage:
2. Generating N particles using uniform distribution for system model: $z_t \in z_{t-1} \pm R$;
3. Updating: weighting each particle;
4. Normalizing the weights;
5. Resampling according to the first round weights;
6. Fine stage:
7. Propagating the resampled N particles using Gaussian model;
8. Updating: re-weighting the N particles;
9. Normalizing the weights;
10. Output the estimation using MAP

3.2 Weight Calculation

In the root, the weight of each particle should reflect the true likelihood with the tracking object’s representation. In our approach, we adopt the partitioned color cue that combines color distribution with shape information implicitly.

The Color Cue. The color cue is based on template matching of color distribution [8,9], where the color histograms are used as the target model.

To make the algorithm robust to lighting conditions, the color space is discretized into m bins and assign each pixel to the corresponding bin. In our experiments, the RGB color value is mapped into $8 + 8 + 8$ bins, i.e. using 8 bins for each color channel.

After obtaining the color histogram for the sample particles and template, the similarity is calculated using the Bhattacharyya distance. Suppose $p = p^{(u)}$, $q = q^{(u)}$, $u = 1 \dots m$ representing the discrete color histograms for a particle p and the template q . The Bhattacharyya coefficient is defined as:

$$\rho[p, q] = \sum_{u=1}^m \sqrt{p^{(u)}q^{(u)}} \tag{8}$$

The more similar of the two distributions are, the larger ρ is. For two identical normalized distributions, the $\rho = 1$. Then, the distance of two distributions is defined as:

$$d[p, q] = \sqrt{1 - \rho[p, q]} \tag{9}$$

Further, in order to integrate the distance in a probabilistic way for the particle filter framework, we generate the particle’s color weight by a Gaussian model:

$$\omega^{(p)} = \frac{1}{\sqrt{2\pi\delta}} \exp^{-\frac{d[p,q]^2}{2\delta^2}} \tag{10}$$

The Gaussian variance δ determines the discrimination power of particles in this cue.

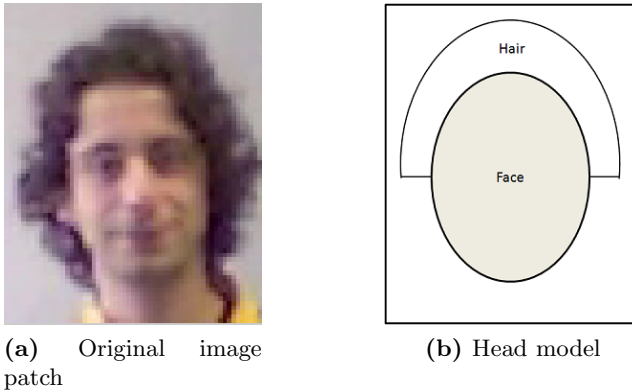


Fig. 1. Partitioned human head model

The Partitioned Color Cue. To complement the color cue, which uses only one distribution representing the whole object and loses information about object's geometric structure, we model the human head with two sections: face and hair, as illustrated in Fig. 1.

Then, we represent the object with 48 bins of color distribution: 24 for the face section and 24 for the hair section.

The tracking target's template can be initialized with a detector algorithm or manually given. In our approach, the initial template is calculated from the first frame specified by its ground truth.

4 Experimental Results

The proposed coarse-to-fine two stages particle filtering algorithm is tested to track the random motion of the human head using a publicly available dataset, SPEVI [16], which consists of 448 frames recorded with a webcam. The video also includes manually marked ground truth so that further quantitative analysis can be performed. The challenge of this video sequence comes from several sources:

- The actor is moving intentionally maneuvering by changing velocity, acceleration, and direction dramatically and randomly;
- The recording device is a low resolution webcam (320 X 240) at the rate of 10 frames per second, some frames are very blurred when fast motion is performed;
- The unstructured background also makes the tracking task difficult not only because the wall area is very similar to the actor's face color, but also the illumination is unevenly changing.

4.1 Results

Fig. 2 illustrates the images captured from the tracking results. As we can observe, the tracking system performs well in different challenging scenarios. For



Fig. 2. Selected frames from the tracking result

example, in the video the images are blurred due to fast movement of the actor (frame 25). The frame series 88-93 even shows that half of the actor's body already moves out of the capture scope of the camera. The appearing size of the actor also varied a lot because of the changing distance to the camera (frame 135 shows small size, whilst frame 36 and 346 show a close-up). Sometimes the actor is present at bright illumination environment (frame 36), and sometimes the actor is at dimmer environment (frame 55 and 295). The tracking system is still able to track most of these situations smoothly.

In order to further evaluate our tracking results quantitatively, the metrics of tracking accuracy is defined as the area overlap between the estimation and ground truth position as [6]:

$$Overlap = \frac{Estimation \cap GroundTruth}{Estimation \cup GroundTruth} \times 100\% \quad (11)$$

Average error for each element of state space, $[x, y, a, b]$, is calculated by average the absolute tracking error in pixels throughout all frames.

In the following each experimental setting, the reported tracking accuracy is averaged by 10 runs and the standard deviation is calculated over all frames within one run.

Time Factor. Fig. 3 plots the tracking accuracy as time proceeds based on 500 particles. From this figure, we can find that as time evolving, the overlap is decreased with several low valleys, which corresponds to the frame number at around 70-100 and 240-300. From the video, we can observe over the two time periods that the actor is in the shadow area where the color of wall background is very similar to the skin color. At the same time, images are also blurred which causes the shape information to be lost too. The tracking accuracy of frame 321 to 325 is zero which correspond to the object of interest is totally going out of the visible range.

The Number of Particles. Next, the factor regarding the needed number of particles is investigated. Fig. 4 plots the average accuracy over 10 runs with

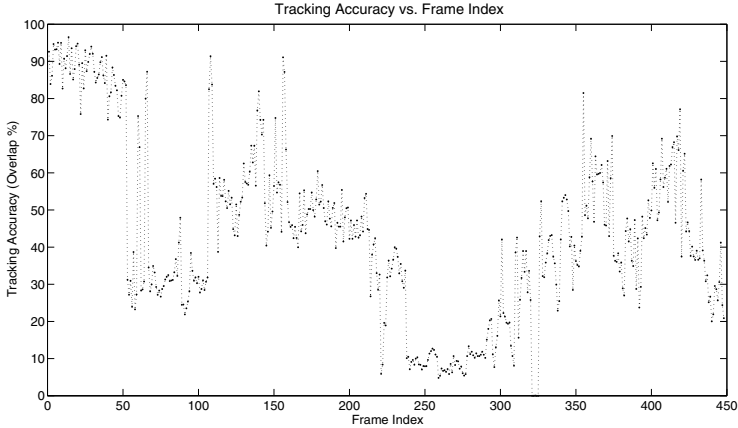


Fig. 3. Tracking accuracy as time proceeds

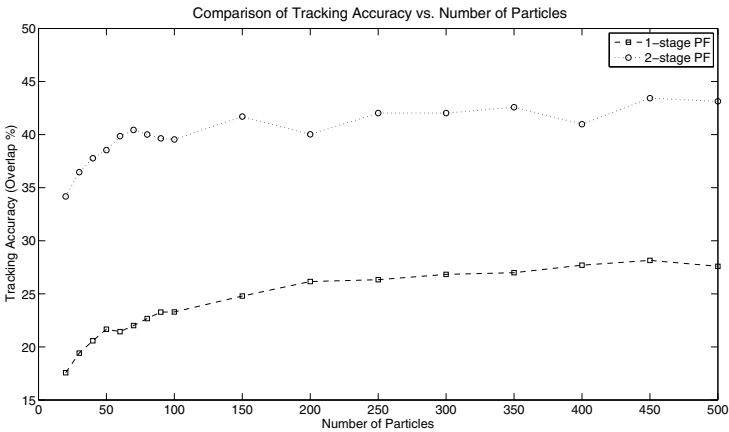


Fig. 4. Comparison of tracking accuracy over the two methods with different number of particles

respect to different number of particles, from 20 to 90 stepped by 10, and from 100 to 500 stepped by 50. To make the two-stage and one-stage PF comparable, the total number of particles for the two-stage method is equivalent to that of the one-stage method.

Generally speaking, the tracking accuracy is improved as the number of particles increases for both coarse-to-fine PF and traditional one-stage PF. For particular low number of particles (below 50), the tracking performance is not well unsurprisingly. Also, when the number of particles reaches above 200, the tracking performance is kept at that level without further improvement. This can be explained as the system is reaching the peak and being limited by the selected cues' representation ability.

Comparing the two methods, observing from the Fig. 4, the two-stage PF achieves high level tracking accuracy with a relative low number of particles (70 and up). This benefit is due to the two rounds of estimation that makes use of samples more efficiently.

Additionally, the tracking results show some fluctuations because the method of particle filter is rooted with randomized generation of samples. However, from Table 1 we can see that the standard deviations of all metrics from the two-stage PF are smaller than that of the one-stage PF. This indicates the coarse-to-fine PF also demonstrates robustness in the same condition.

Table 1. Performance comparison between one-stage and two-stage PF (500 particles)

	Overlap	Error_x	Error_y	Error_a	Error_b
1-Stage	28.54 ± 24.58	41.24 ± 32.48	36.85 ± 44.43	11.99 ± 8.78	15.69 ± 10.84
2-Stage	43.70 ± 23.93	12.11 ± 14.13	15.03 ± 15.50	9.70 ± 7.80	11.88 ± 9.43

5 Conclusion

In this paper, we investigated the problem of tracking maneuvering motion of human head. The motion model of a maneuvering object cannot be formulated by a simple linear form. Therefore, a two-stage particle filter is proposed to enhance the system. In the coarse stage, the particles adopt uniform distribution which is parameterized by the limitation of its motion range within each time step. In the following fine stage, the particles are resampled using the result of the coarse stage. The system is tested with a challenging video sequence for tracking a varied moving human head. The results demonstrate that the two-stage approach is capable of handling maneuvering motion more accurately and robustly through a relatively small number of particles.

To extend this work, we plan to investigate some cues that are robust to illumination changes, such as texture.

Acknowledgments. This work has been supported through the Collaborative Research and Development (CRD) project “DIScrimination of Critical Objects and EVents in PErvasive Multimodal SuRveillance Systems (DISCOVER)” funded by Natural Sciences and Engineering Research Council of Canada (NSERC) and Thales Canada Inc.

References

1. Arulampalam, M., Maskell, S., Gordon, N., Clapp, T.: A tutorial on particle filters for online nonlinear/non-gaussian bayesian tracking. *IEEE Transactions on Signal Processing* 50, 174–188 (2002)
2. Fieguth, P.: *Statistical Image Processing and Multidimensional Modeling*, ch. 4, pp. 85–127. Springer, Heidelberg (2010)
3. Hol, J., Schon, T., Gustafsson, F.: On resampling algorithms for particle filters. In: *Nonlinear Statistical Signal Processing Workshop*, pp. 79–82 (2006)

4. Houtekamer, P., Mitchell, H.L.: Data assimilation using an ensemble kalman filter technique. *Monthly Weather Review* 126, 796–811 (1998)
5. Julier, S.J., Uhlmann, J.K.: Unscented filtering and nonlinear estimation. *IEEE Review* 92(3), 401–422 (2004)
6. Kasturi, R., Goldgof, D., Soundararajan, P., Manohar, V., Garofolo, J., Bowers, R., Boonstra, M., Korzhova, V., Zhang, J.: Framework for Performance Evaluation of Face, Text, and Vehicle Detection and Tracking in Video: Data, Metrics, and Protocol. *IEEE Trans. Pattern Analysis and Machine Intelligence* 31(2), 319–336 (2009)
7. Li, Z., Kulić, D.: Particle filter based human motion tracking. In: *IEEE International Conference on Control, Automation, Robotics and Vision 2010* (2010)
8. Nummiaro, K., Koller-Meier, E., Gool, L.V.: An adaptive color-based particle filter. *Image and Vision Computing* 21(1), 99–110 (2003)
9. Pérez, P., Hue, C., Vermaak, J., Gangnet, M.: Color-based probabilistic tracking. In: Heyden, A., Sparr, G., Nielsen, M., Johansen, P. (eds.) *ECCV 2002*. LNCS, vol. 2350, pp. 661–675. Springer, Heidelberg (2002)
10. Shafique, K., Shah, M.: A non-iterative greedy algorithm for multi-frame point correspondence. In: *Proceeding of IEEE International Conference on Computer Vision (ICCV)*, pp. 110–115 (2003)
11. Shen, C., Hengel, A.v.d., Dick, A.: Probabilistic multiple cue intergration for particle filter based tracking. In: *Proceeding of VIIth Digital Image Computing: Techniques and Applications*, pp. 399–408 (2003)
12. Sung, H., Choi, K., Cho, S., Byun, H.: Coarse-to-fine particle filter by implicit motion estimation for 3d head tracking on mobile devices. In: *20th International Conference on Pattern Recognition (ICPR 2010)*, pp. 3615–3618 (2010)
13. Veenman, C., Reinders, M., Backer, E.: Resolving motion correspondence for densely moving points. *IEEE Trans. Patt. Analy. Mach. Intell.* 23(1), 54–72 (2001)
14. Welch, G., Bishop, G.: An introduction to the kalman filter. Technical Report, TR95-041, Computer Science, UNC Chapel Hill (1995)
15. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. *ACM Comput. Surv.* 38(4), 1–45 (2006)
16. Surveillance performance evaluation initiative (spevi) dataset, <http://www.eecs.qmul.ac.uk/~andrea/spevi.html> (last accessed: January 2011)

Multi-camera Relay Tracker Utilizing Color-Based Particle Filtering

Xiaochen Dai and Shahram Payandeh

Experimental Robotics and Graphics Laboratory,
School of Engineering Science, Simon Fraser University,
8888 University Drive, Burnaby, BC, V3J 2W4, Canada.

xda6@sfu.ca, shahram@cs.sfu.ca

Abstract. This paper presents a multi-camera surveillance system for motion detection and object tracking based on Motion History Image (MHI), Color-based Particle Filtering (CPF), and a novel relay strategy. The system is composed of two Pan-Tilt-Zoom (PTZ) cameras completely calibrated and placed on desks. Initially, both cameras work as stationary Scene View Camera (SVC) to detect objects for abnormal human motion events such as sudden falling using MHI. If an object is detected in one camera, the other camera can then be controlled to work as Object View Camera (OVC), follow this object, and get zoom-in images using CPF. The states of the tracked object can be exchanged across cameras so that in case that the OVC loses the object, the SVC has sufficient knowledge of the object location, and it can become a new OVC to run the tracking relay. Meanwhile, the original OVC should be reset to work as SVC in order not to lose the global view. Two scenarios, in which the cameras have large or little overlapping field of view, are proposed and analyzed. Experimental study further demonstrates the effectiveness of the proposed system.

Keywords: multi-camera, MHI, CPF, relay strategy.

1 Introduction

Visual detection and tracking is one of the active and challenging research topics in machine vision. Several approaches have been proposed previously. The mean shift algorithm is used but it requires the entire object patch be visible and does not work properly in cases of occlusions or abrupt motions [1, 2]. The Kalman filter, as a probabilistic prediction tool, is more robust to track objects under occlusions [3, 4, 5]. However, the motion should be correctly modelled and the posterior density function (PDF) is strictly assumed to be Gaussian. Compared to the Kalman filter, the particle filtering is a sequential Monte Carlo algorithm that does not assume specific type of densities and has the ability to handle occlusions as well. For example, Nummiaro et al. [6] propose an algorithm to add an adaptive appearance model based on color distributions to particle filtering.

With the decreasing cost of image sensors and the increasing computational capability of supporting processors, the potential to include multiple cameras

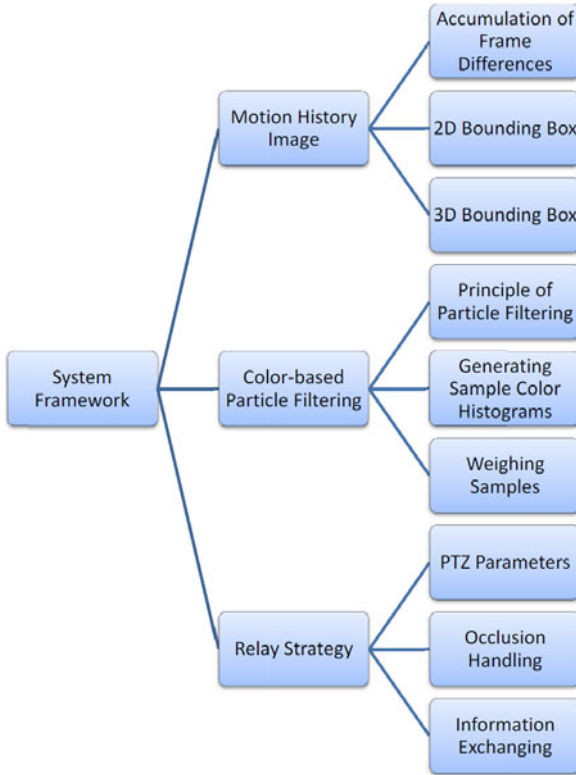


Fig. 1. System Framework

has become more feasible [7]. Multiple cameras can enhance the total coverage of a scene and recover 3D depth of the objects. Khan et al. [8] and Javed et al. [9] use single camera tracking results along with the relation between boundaries of camera field of view (FOV) to establish correspondence between views of the same object in multiple cameras. Moreover, the features of Pan-Tilt-Zoom (PTZ) cameras not only expand the camera view through panning and tilting, but also direct attention to details through zooming [10]. Mottaghi et al. [11] develop an optimized tracking approach using multiple PTZ cameras and Lu et al. [12] build up a cooperative hybrid multi-camera tracking system.

In this paper, we propose a surveillance system consisting of two calibrated cameras which can work as either stationary Scene View Camera (SVC) or moving Object View Camera (OVC). Initially, both cameras work as SVC and have a wide view of the scene. Motion History Image (MHI) [15] is utilized to detect events of interest and 3D projected bounding boxes are generated to represent moving objects using the approach we propose in [13]. Since camera placement plays a critical role in the process of event detection, in our experiment we put the cameras on ordinary desks to investigate the motion of sudden falling based on the dynamic states of the 3D bounding boxes. Two scenarios are designed

for the experimental study. In the first scenario, the two cameras have a large overlapping FOV and one camera may trigger the other to work as OVC if a fall is detected. In the second scenario, the two cameras have little overlapping FOV and states of the object can be exchanged across cameras to form a tracking relay.

Sensitivity analysis of camera calibration has been studied to obtain better calibration result [14], which is a critical factor for controlling the OVC. With an accurate camera calibration, the position of the object of interest is obtainable so that the OVC knows where to pan and tilt in order to fit the object in its central view, and then zoom in to obtain various levels of detail. Color-based particle filter (CPF) is implemented for object tracking. We use Hue-Saturation-Value (HSV) color space model to generate multi-dimensional color histograms, and then weigh samples through particle filtering. The novelty of the proposed approach mainly lies in the mixture of PTZ capability, event detection, and relay strategy about exchanging information across cameras that all together yield a reliable system. Fig. 1 shows a framework of the proposed system.

The outline of the rest of this paper is as follows. In Section II, we state motion event detection using MHI in SVC. In Section III, we present the CPF for tracking people in OVC. Section IV explains the relay strategy for exchanging information across cameras. Experimental results for our visual tracking system are provided in Section V followed by concluding remarks in Section VI.

2 Motion Event Detection Using MHI

MHI is used to detect human motion events such as sudden falling in SVC. Our novel contribution of 3D projected bounding boxes are developed to represent silhouettes of moving objects. These 3D bounding boxes are generated based on the multiple view geometry. Compared to traditional 2D bounding boxes, they use one more dimension to reflect the dynamic status of objects, and thus can be applied for the analysis of human motions. The specific procedure to generate a 3D box is proposed in [13].

Stationary SVC is used to detect human motion events. To recover a series of motions, we generate MHI for a sequence of $N + 1$ successive frames with N frame differences to be obtained and the accumulation of frame differences can yield directional motion information of objects. We find some limitations if only single camera is used. For example, if the object is close to the center of the camera view or the motion occurs along the optical axis, very little changes of the boundary can be detected. Hence, the camera fails in the detection. This problem can be addressed by introducing multiple cameras that are positioned appropriately.

Moving objects in SVC are first represented by individual bounding boxes whose centers and sizes are calculated according to the N frame differences layered in the MHI. We weight the successive frame differences and calculate the center, width, and height of the bounding box through weighted sum of all the sub-bounding boxes generated from each pair of neighbouring frame differences.

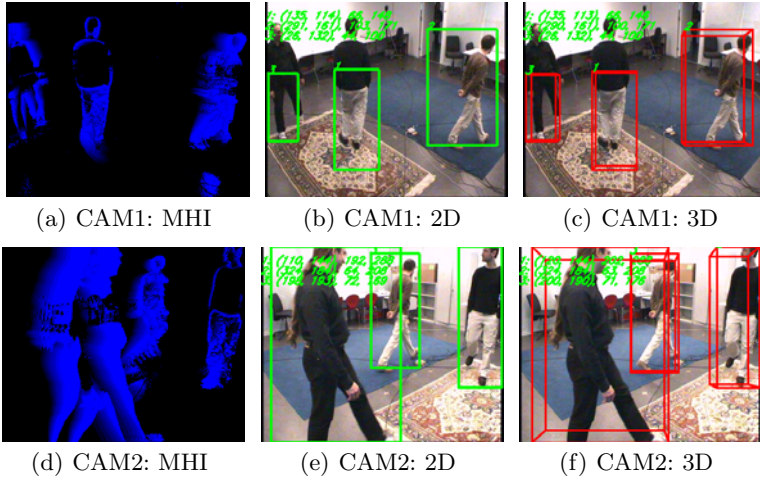


Fig. 2. Two cameras work as SVC to detect human motions in a room

Fig. 2 shows results of the proposed approach using the CVLAB dataset [16]. The left column shows that three persons are walking around in a room and MHI ($N = 5$) is applied to detect their motions in two SVCs at different locations. 2D bounding boxes are tagged, and their center, width, and height are computed, as shown in the middle column. The multi-camera setting-up makes it possible to obtain spatial geometrical information of the objects. Based on the results of camera calibration, In the right column, we construct projected 3D bounding boxes to represent the walking persons.

A typical motion event, falling, is defined and experimented for this system. The judging criterion is based on the dynamic states of the bounding boxes. Fall happens when the height of a bounding box at time t , h_t and vertical location, v_t dramatically decrease, which means $h_t = \frac{h_t - h_{t-1}}{h_{t-1}} < \Delta h$ and $v_t = \frac{v_t - v_{t-1}}{h_{t-1}} < \Delta v$. h'_t and v'_t are the instantaneous changes of height and vertical position of the bounding box. Δh and Δv are thresholds defined during experiment.

3 Adaptive Color-Based Particle Filtering

The particle filtering is a sequential Monte Carlo method that requires a multi-variate Gaussian distribution and makes few assumptions on either the transition model or the sensor model. An overview of the particle filtering for visual tracking can be found in [17]. The general idea of particle filtering is to find the posterior density function (PDF)

$$p(x_t | z_1, \dots, z_t) = \frac{p(x_t | z_1, \dots, z_{t-1}) p(z_t | x_t, z_1, \dots, z_{t-1})}{p(z_t | z_1, \dots, z_{t-1})} \quad (1)$$

where x_t and z_t are state and measurement at time t , respectively. Since z_t is independent of z_{t-1} , according to Bayes' rule, $p(z_t | x_t, z_1, \dots, z_{t-1}) = p(z_t | x_t)$.

Also $p(z_t|z_1, \dots, z_{t-1})$ is a constant. We have

$$p(x_t|z_1, \dots, z_t) = kp(z_t|x_t)p(x_t|z_1, \dots, z_{t-1}) \quad (2)$$

where k is a normalization factor that does not depend on x_t . We use the recursive definition to compute the filtered distribution $p(x_t|z_1, \dots, z_t)$ given the distribution $p(x_{t-1}|z_1, \dots, z_{t-1})$. With a particle representing $p(x_t|z_1, \dots, z_{t-1})$, we may create a dynamic model to approximate x_t as

$$x_t = f(x_{t-1}, n_{t-1}) \quad (3)$$

where f is a non-linear function of the previous state x_{t-1} and n_{t-1} is a sequence. Given a set of N samples S , such that

$$S = \left\{ \left(s_t^{(i)}, \pi_t^{(i)} \right) \mid i = 1, \dots, N \right\} \quad (4)$$

Each sample is given a weight as

$$\pi_t^{(i)} = p(z_t|x_t = s_t^{(i)}); \left(\sum_{i=1}^N \pi_t^{(i)} = 1 \right) \quad (5)$$

Therefore, the estimated expectation of state vector at time t is

$$E(x_t) = \sum_{i=1}^N \pi_t^{(i)} s_t^{(i)} \quad (6)$$

As the number of samples grows, particle filter can recover true PDF, and thus provides a robust tracking framework. For example, if a tracked person is partially or completely occluded for tens of frames, the tracker can still retrieve correct tracking when the person reappears.

Once a motion event is detected and the object is centered in the view of OVC, CPF is used to track the object. A color histogram generated based on the color distribution within the object bounding box is created as a reference to weigh samples in the particle filter. We use the color histogram created from the Hue-Saturation-Value (HSV) color space model, which separates the chromatic information of hue, saturation and value from the intensity. Generally, a color histogram with more dimensions makes the CPF more robust to color distraction in the background, but it also results in the exponentially increased computational cost. Compared with [12] in which 2D histogram composed of $m_h m_s$ bins is used, where m_h and m_s are the numbers of hue and saturation bins used, we use the color histogram consisting of $m = m_h m_s + m_v$ bins, where the extra m_v is the number of value bins used and this proves to enhance the robustness if there exists varying illumination. The normalized histogram may represent the probability of hue, saturation and value that the object has. The object to be tracked is represented by a 2D bounding box whose state vector s_t is a 9-tuple vector

$$s_t = \left\{ u_t, v_t, u'_t, v'_t, w_t, h_t, w'_t, h'_t, s \right\} \quad (7)$$

(u_t, v_t) is the center of the bounding box at time t . u'_t and v'_t are the instantaneous velocity of the box moving in the directions of axes u and v in image the image frame, respectively, at time t . w_t and h_t are the width and height of the box at time t . w'_t and h'_t are the instantaneous changes of the width and height at time t , s is the scaling factor of each sample.

To weigh the sample sets, we collect the color histogram of the target H_T and the color histogram of the sample H_S^i . Due to the factor that boundary pixels of the bounding box usually belong to the background, we assign smaller weights to those pixels that are further away from the center of camera view by employing a weighing function

$$k_r = \begin{cases} 1 - \frac{4r^2}{u_t^2 + v_t^2} & r < \frac{1}{2}\sqrt{u_t^2 + v_t^2} \\ 0 & \text{otherwise} \end{cases} \quad (8)$$

where r is the distance between the corresponding pixel and the center pixel within the bounding box. The Bhattacharyya distance d_B , which measures the similarity between two discrete probability distributions H_T and H_S^i , is computed as

$$d_B = \sqrt{1 - \sum_{j=1}^m \sqrt{k_r H_S^i(j) H_T(j)}} \quad (9)$$

The smaller the value of d_B , the similar the samples to the target. Smaller d_B corresponds to large weights. The weight function defined in Equation (5) is chosen to be

$$\pi_t^{(i)} = \frac{1}{\sqrt{2\pi_i}\sigma_i} e^{-\frac{d_B^2}{2\sigma_i^2}} = \frac{1}{\sqrt{2\pi_i}\sigma_i} e^{-\lambda d_B^2} \quad (10)$$

where λ is scaling factor determined through experiments. The state of the bounding box at $t = 0$ should be set to initiate the CPF based on the information from SVC.

4 Relay Strategy for Exchanging Information

In a multi-camera system, collaboration between each pair of cameras is important. Firstly, if a motion event is detected in one camera, we try to localize it and initialize the CPF tracking in the other camera based on camera geometry. Secondly, in case one of the cameras loses the target, the other camera can provide reliable information where to look in order to retrieve the target.

As shown in Fig. 3, We suppose that an object point P in space is projected onto the image plane of one camera (CAM1) at p_1 and the image plane of the other camera (CAM2) at p_2 . The key to localize the target is to compute the coordinate of $p_2 = (u_2, v_2)$, if given the coordinate of $p_1 = (u_1, v_1)$ and calibration parameters. (X_{c1}, Y_{c1}, Z_{c1}) and (X_{c2}, Y_{c2}, Z_{c2}) are the 3D coordinates of P in the camera frames of CAM1 and CAM2, respectively, and their values are obtainable. Details of derivation can be found in [12].

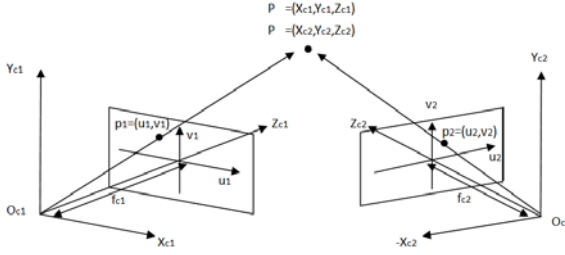


Fig. 3. Geometry relationship of two cameras

If CAM2 needs to rotate about Y_{c2} (panning) for α degrees and then rotate about the new X_{c2} (tilting) for β degrees in order to localize the target on its view center, the rotation angles can be calculated as

$$\alpha = -\arctan \frac{X_{c2}}{Z_{c2}}; \beta = \arctan \frac{Y_{c2}}{Z_{c2} \cos \alpha - X_{c2} \sin \alpha} \quad (11)$$

The OVC can zoom in to obtain close images of the target with the zoom-in amount $\delta_f = f'_2 - f_2$, where f_2 is the original focal length and f'_2 is the focal length after zoom-in. The value of f'_2 is computed as

$$f'_2 = \frac{Z'_2 a_r H}{H^m m_{v2}} \quad (12)$$

where

$$Z'_2 = -X_{c2} \sin \alpha \cos \beta + Y_{c2} \sin \beta + Z_{c2} \cos \alpha \cos \beta \quad (13)$$

is the projection of target along the z-axis of camera frame before zooming in; H^m is the height of the person in meters; m_{v2} is the number of pixels per unit distance along y-axis.

The relay strategy for exchanging object state information across cameras is implemented in the situation when the tracked object runs out of the view of OVC due to reasons such as occlusion. In these cases, the SVC is triggered to work as a new OVC and the original OVC is reset to be a SVC so that the system still keeps monitoring the global environment.

To initialize CPF in the new OVC, the object state vector at the time of target handing over $s_{t1} = \{u_{t1}, v_{t1}, u'_{t1}, v'_{t1}, w_{t1}, h_{t1}, w'_{t1}, h'_{t1}, s_{t1}\}$ is set as follows. u_{t1} and v_{t1} are set as the image center since the CPF tracking in OVC starts after panning and tilting to fit the target on the center view. u'_{t1} and v'_{t1} are the estimated velocity of the object. For a person walking at a normal speed of $1.0m/s^2$, u'_{t1} is set as $f m_u / Z'_{c2}$ and v'_{t1} is set as $f m_v / Z'_{c2}$, where f is the focal length, m_u and m_v are the numbers of pixels per unit distance along x-axis and y-axis, respectively, and Z'_{c2} is defined in Equation (13). w_{t1} and h_{t1} are the dimension of the bounding box achieved by computing the projected bounding box between different views. w'_{t1} , h'_{t1} and s_{t1} are set flexibly in experiment. In our system, we set $w'_{t1} = h'_{t1} = 0$ and $s_{t1} = 1$.

5 Experimental Results

Experimental results demonstrate the performance of our tracking system. We use two cameras (Sony EVI-D100) with PTZ capabilities to capture images having 640×480 resolution from live video streams at a frame rate of 30 fps. Both

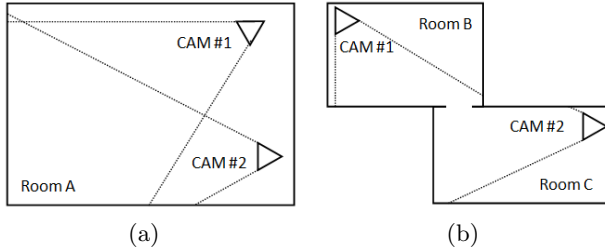


Fig. 4. Camera set-up and ground plan of two scenarios. The dash lines indicate the FOV lines.

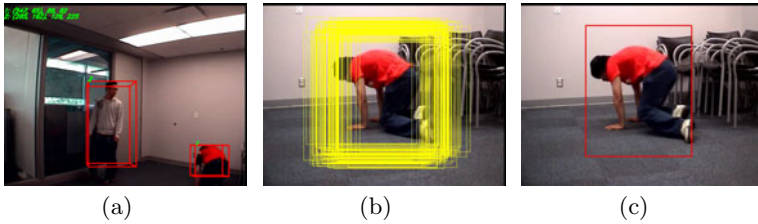


Fig. 5. Scenario One: (a) A fall is detected in CAM 1 as SVC; (b) Samples ($N=100$) propagated in CAM 2 as OVC; (c) Weighed samples

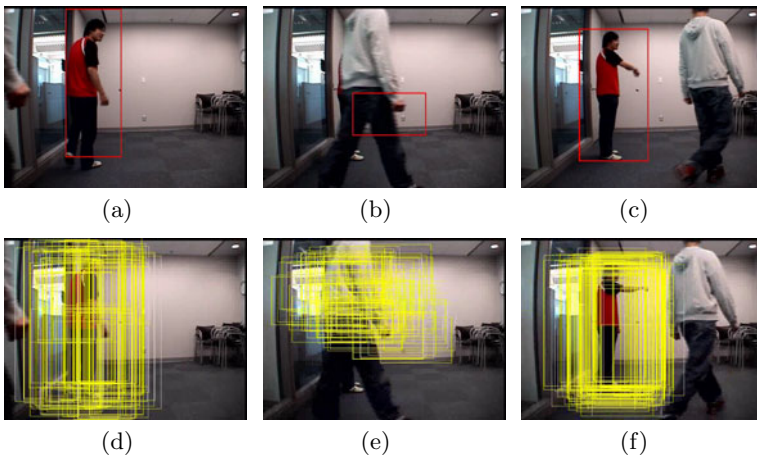


Fig. 6. Occlusion handling by using CPF in OVC

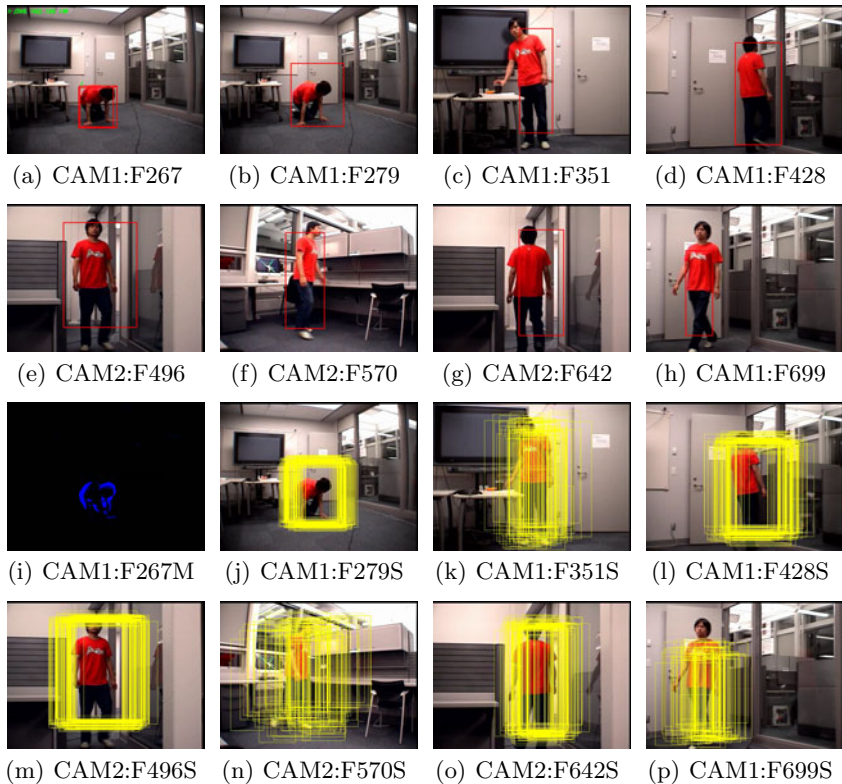


Fig. 7. Scenario Two: Snapshots captured by two cameras localized in two connected rooms. The name of sub-figures is defined as F+frame number+(M: MHI or S: samples). (a)-(d) and (h) are captured by CAM1 with weighted bounding boxes; (e)-(g) are captured by CAM2 with weighted bounding boxes; (i) MHI of (a); (j)-(p) propagated samples corresponding to (b)-(h), respectively.

cameras are attached to the same computer for data processing. Two scenarios have been composed. The camera set-up and ground plan is shown in Fig. 4.

In the first scenario as shown in Fig. 4(a), two cameras are located in a single room, sharing a large overlapping field of view. Two persons are walking around in the room. Initially, MHI is utilized for the SVC to detect human motions and 3D projected bounding boxes are generated, encapsulating the persons. One of the cameras detects a fall event happening to a person, and then triggers the other camera to work as OVC. The OVC then pan and tilt to fit the person to the center of its view, and track him using CPF. Meanwhile, it zooms in to obtain large and clear images of the object. The Threshold for fall detection is set as $\Delta_h = -0.50$ and $\Delta_v = -0.25$ during the experiment so as to differentiate it from other normal motions. Fig. 5 gives the snapshots when one SVC detects the fall event and then triggers the other camera to work as OVC. The number of samples is $N = 100$.

In Fig. 6, the top row images use a single bounding box weighted by propagated samples from the corresponding bottom row images. We find that the proposed CPF method is capable of handling complete occlusion which lasts for tens of frames as can be seen in the middle column images, where the person being tracked is blocked by a newcomer.

In the second scenario as shown in Fig. 4(b), two cameras are localized in two connected rooms sharing very little overlapping field of view. Fig. 7 are the snapshots of the two cameras from video streams containing around 700 frames. At the beginning, both cameras work as SVC, heading to the door which connects these two rooms. As an individual person in Room B falls onto the ground, CAM1 is triggered and works as OVC to keep tracking the person. This is shown in Fig. 7(a)-(d). The person then leaves Room B and enters Room C. Since CAM1 completely loses the object, State vector of CPF for the person is initialized in CAM2, which gets the relay to track the person. Finally, the person returns to Room B from Room C as displayed in Fig. 7 (p), and CAM1 successfully retrieves the correct tracking.

6 Discussion and Future Work

In this paper, CPF and a relay strategy are utilized in a visual surveillance system for real-time event detection and object tracking. MHI is used in the SVC to detect human abnormal motions such as falling and 3D projected bounding boxes are generated to represent the objects. SVC can be controlled to change into OVC and then it can track the object by panning, tilting, and zooming in. Multi-dimensional color histogram based on the HSV color space model is fused with the particle filtering method to achieve accuracy people tracking. Compared with some of the previous work, the main contribution of our work is the integration of PTZ capability, event detection, and relay strategy that all together yield an intelligent and reliable system.

One focus of our future work will be to investigate into detail technical problems about the relation between CPF and the zooming process. For example, how the zooming process affects the particles of the filter and how the uncertainty in the target position affects the focal length of the PTZ camera? Besides, if 3D bounding box is used in OVC to represent the traced object, depth information may also be included in the state vector s_t in Equation (7). In addition, we may combine more features (e.g., gradient orientation distribution) with the color distribution in the particle filter because the sole color likelihood does not seem to be sufficient to distinguish the target occluded by objects with the similar color. Another potential direction will be to study more crowded scenes, possibly by using more cameras so that we may have more sensors available to monitor not only the individual object, but the global scene as well to handle events occurring at any time.

References

1. Comaniciu, D., Ramesh, V., Meer, P.: Real-time Tracking of Non-rigid Objects Using Mean Shift. In: International Conference on Computer Vision and Pattern Recognition, vol. 2, pp. 142–149 (2000)
2. Shabani, A., Ghaemini, M., Shokouhi, S.B.: Human Tracking Using Spatialized Multi-level Histogram and Mean Shift. In: Canadian Conference on Computer and Robot Vision, pp. 151–158 (2010)
3. Wang, J., He, F., Zhang, X., Gao, Y.: Tracking Objects through Occlusions Using Improved Kalman Filter. In: 2nd International Conference on Advanced Computer Control, vol. 5, pp. 223–228 (2010)
4. Mehta, M., Goyal, C., Srivastava, M.C., Jain, R.C.: Real Time Object Detection and Tracking: Histogram Matching and Kalman Filter Approach. In: International Conference on Computer and Automation Engineering, vol. 5, pp. 796–801 (2010)
5. Wang, Y., Liu, T., Li, M.: Object Tracking with Appearance-based Kalman Particle Filter in Presence of Occlusions. In: WRI Global Congress on Intelligent Systems, vol. 1, pp. 288–293 (2009)
6. Nummiaro, K., Koller-Meier, E., Gool, L.V.: An Adaptive Color-based Particle Filter. *Image and Vision Computing* 21(1), 99–110 (2002)
7. Aghajan, H., Cavallaro, A.: *Multi-camera Networks Principles and Applications*. Academic Press, London (2009)
8. Khan, S., Javed, O., Rasheed, Z., Shah, M.: Human Tracking in Multiple Cameras. In: International Conference on Computer Vision, vol. 1, pp. 331–336 (2001)
9. Javed, O., Rasheed, Z., Alatas, O., Shah, M.: KNIGHT: A Real Time Surveillance System for Multiple and Non-overlapping Cameras. In: International Conference on Multimedia and Expo, vol. 1, pp. 649–652 (2003)
10. Del Bimbo, A., Dini, F., Grifoni, A., Pernici, F.: Exploiting Single View Geometry in Pan-Tilt-Zoom Camera Networks. In: International Workshop on Multi-camera and Multi-modal Sensor Fusion (2008)
11. Mottaghi, R., Payandeh, S.: Coordination of Multiple Agents for Probabilistic Object Tracking. In: Canadian Conference on Computer and Robot Vision, pp. 162–167 (2005)
12. Lu, Y., Payandeh, S.: Cooperative Hybrid Multi-camera Tracking for People Surveillance. *Canadian Journal of Electrical and Computer Engineering* 33(3), 145–152 (2008)
13. Dai, X., Payandeh, S.: Toward Spatial Tracking in Multiple Camera Environment. In: Proceedings of IEEE Canadian Conference on Electrical and Computer Engineering (2010)
14. Lu, Y., Payandeh, S.: On the Sensitivity Analysis of Camera Calibration from Images of Spheres. *Journal of Computer Vision and Image Understanding* 114(1), 8–20 (2009)
15. Davis, J.W.: Hierarchical Motion History Image for Recognizing Human Motion. In: Proceedings of IEEE Workshop on Detection and Recognition of Events in Video (2001)
16. Fleuret, F., Berclaz, J., Lengagne, R., Fua, P.: Multi-Camera People Tracking with a Probabilistic Occupancy Map. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 30(2), 267–282 (2008)
17. Mottaghi, R., Payandeh, S.: An Overview of a Probabilistic Tracker for Multiple Cooperative Tracking Agents. In: Proceedings of 12th IEEE International Conference on Advanced Robotics, pp. 888–894 (2005)

Visual Tracking Using Online Semi-supervised Learning

Meng Gao¹, Huaping Liu^{2,3}, and Fuchun Sun^{2,3}

¹ Shijiazhuang Tiedao University, Shijiazhuang, Hebei Province, P.R.China

² Department of Computer Science and Technology, Tsinghua University, P.R.China

³ State Key Laboratory of Intelligent Technology and Systems, Beijing, P.R.China

Abstract. Since there does not exist labelled samples during tracking period, most existing classification-based tracking approaches utilize a “self-learning” to online update the classifier. This often results in drift problems. Recently, semi-supervised learning attracts a lot of attentions and is incorporated into the tracking framework which collects unlabelled samples and use them to enhance the robustness of the classifier. In this paper, we develop a gradient semi-supervised learning approaches for this application. During the tracking period, the semi-supervised technology is used to online update the classifier. Experimental evaluations demonstrate the effectiveness of the proposed approach.

Keywords: Visual tracking, semi-supervised learning.

1 Introduction

Object tracking is an important problem with extensive applications in domains including video surveillance, robot guidance and human-computer interaction [7] and therefore attracted significant interest in the compute vision community [15]. The goal of visual tracking is to automatically locate the same object in adjacent frames in a video sequence.

Since tracking is a time-dependent problem, an adaptive mechanism is more suitable for this application. [3] firstly proposed a method to adaptively select color features that best discriminate the object from the current background. Another important work is [1], which used an adaptive ensemble of classifier. However, [1] does not update the weak classifiers themselves, but replaces some of the older weak classifiers with new weak classifiers. [5] designed an on-line boosting classifier that selects features to discriminate the object from the background. It models the feature density by simple Gaussian and update their parameters using Kalman filter with some specified parameters. However, the Gaussian model is not sufficient to characterize background samples. In [16], a classifier adaption approach is proposed to improve an existing generic classifier. The main idea is that the cost function on the old and new training data-set are combined in a compact Taylor expansion form. However, the parameter controlling the relative importance of the old and new data-set is difficult to determine. In addition, the

approach proposed in [16] was used for adjusting voting weights of weak classifiers, but not classifiers themselves. Very recently, [12] and [9] proposed two online boosting approaches which can adaptively adjust classifier parameters. In [9], a gradient-based feature selection approach is used and show promising results in object tracking and classifier updating.

These “classification-based tracking” approaches are so promising that many scholars combined them with the popular particle filter. For example, [2] used the approach proposed in [3] to design a particle filter with adaptive feature selection ability. [14] embedded the feature selection procedure into the particle filter with the aid of existing “background” particles. [8] proposed a cascaded particle filter with discriminative observers of different lifespan.

However, it should be noticed that all the above-mentioned classifiers are online updated in a “self-learning” manner, i.e., the estimated target region is used to extract new positive samples and the surrounding regions are used to extract negative samples. In practice these “positive” or “negative” samples are not reliably labelled since the “teacher” is just the current tracking results. Once minor bias occurs during tracking period (This is very often in tracking applications), the assigned labels may be noisy. Therefore these “self-learning” approaches usually tend to “drift” since the error may be accumulated during the learning and tracking process. In fact, in many tracking applications, the labelled samples are given by an extra detector which only works at the first frame and therefore the number of labelled samples is very small, while the unlabelled samples, which can be selected from any frame, is enormous and easy to get. If we wish to update the classifier online, we should not ignore the unlabelled samples. This motivates us to use the popular semi-supervised learning approach [17].

Though the semi-supervised learning achieves great successes, its application in tracking domain is still very rare. Recently, [13] utilized the co-training SVM approach to design a semi-supervised tracker. A demerit of this approach is that the tracker needs several initial frames to get enough labelled samples. In tracking scenarios, extracting feature from the first frame only is more attractive. In addition, the co-training approach requires calculating different visual cues. In [6], a SemiBoost-based online boosting approach was used for tracking, which is a straightforward extension of the supervised online boosting approach [5]. To avoid some intrinsic problems in SemiBoost, [5] just developed a very simplified SemiBoost version. In this paper, we propose a semi-supervised tracking approach. Different from existing works, we use the unlabelled samples extracted during tracking period to improve existing classifier, but not construct new classifier. The updating procedure is based on gradient descent and therefore temporal gradient learning is coupled with the intrinsic gradient learning in Boosting-like approaches. By using this class of gradient learning, the information stored in previous classifiers are kept and only some necessary modifications on the classifiers will be made.

2 Gradient Semi-supervised Learning

Consider a dataset $\{\mathbf{f}_1, \mathbf{f}_2, \dots, \mathbf{f}_n\}$ (the last rows of all samples are 1 that is used for calculating the intercept.) and the corresponding label $\{y_1, y_2, \dots, y_n\}$. The label y_i takes value from the set $\{+1, 0, -1\}$, where $+1$, -1 and 0 represent positive, negative and unlabelled labels, respectively. For conveniences, we denote \mathcal{F}_l as the index set of labelled samples, and \mathcal{F}_u as the index set of unlabelled samples, i.e., $\mathcal{F}_l = \{i|y_i \neq 0\}$ and $\mathcal{F}_u = \{i|y_i = 0\}$. The goal of semi-supervised learning is to use the labelled samples and unlabelled samples to construct a robust classifier.

Algorithm 1. GentleBoost algorithm

Given: Labelled samples $\{\mathbf{f}_i\}_{i \in \mathcal{F}_l}$, label $\{y_i\}_{i \in \mathcal{F}_l}$, iteration number T .

OUTPUT: $H(\mathbf{f}) = \sum_{t=1}^T h_t(\mathbf{f})$

Initialize: $H(\mathbf{f}) = 0$, $w_i = 1$ for all $i \in \mathcal{F}_l$

FOR $t = 1, 2, \dots, T$

– Determine the parameter of weak classifier as $\beta_t = \operatorname{argmin}_{\beta} \{\sum_{i \in \mathcal{F}_l} w_i (\beta^T \mathbf{f}_i - y_i)^2\}$.

– Define $h_t(\mathbf{f}) = \frac{2}{\pi} \operatorname{atan}(\beta_t^T \mathbf{f})$.

– Compute the weight for $i \in \mathcal{F}_l$: $w_i = w_i e^{-y_i h_t(\mathbf{f}_i)}$ and normalize it to satisfy $\sum_{i \in \mathcal{F}_l} w_i = 1$.

– Update the classifier as $H(\mathbf{f}) = H(\mathbf{f}) + h_t(\mathbf{f})$.

The main idea of the proposed gradient semi-supervised learning to design a preliminary classifier using labelled sample set \mathcal{F}_l only, and then use the unlabelled sample set \mathcal{F}_u to improve its performance. Therefore we should first utilize conventional Adaboost algorithm to construct a preliminary classifier. In this paper, we select GentleBoost for further conveniences. In this setting, $H(\cdot)$ is an ensemble classifier which can be specifically represented as

$$H(\mathbf{f}) = \sum_{t=1}^T h_t(\mathbf{f}) \quad (1)$$

where $h_t(\cdot)$ is a weak classifier which takes value from the continuous interval $[-1, +1]$, and T is the number of weak classifiers. For any sample \mathbf{f} , the practical classifier output is $\operatorname{sign}(H(\mathbf{f}))$.

Given sample vector \mathbf{f}_i , the t -th weak classifier is designed as

$$h_t(\mathbf{f}) = \frac{2}{\pi} \operatorname{atan}(\beta_t^T \mathbf{f}) \quad (2)$$

where β_t , the parameter vector for $h_t(\cdot)$, can be calculated by weighted least square approach. For more details, please refer to [9] and [12].

Algorithm 2. Gradient learning algorithm

Given: Unlabelled samples $\{\mathbf{f}_i\}_{i \in \mathcal{F}_u}$, Initial parameter $\{\beta_t\}_{t=1}^T$
 OUTPUT: Updated parameter $\{\tilde{\beta}_t\}_{t=1}^T$

- FOR $t_0 = 1, 2, \dots, T$
 - $\tilde{\beta}_{t_0} = \beta_{t_0}$
 - **While (1)**
 - Form the current classifier according as

$$H(\mathbf{f}) = \frac{2}{\pi} \left\{ \sum_{t=1}^{t_0-1} \text{atan}(\tilde{\beta}_t^T \mathbf{f}) + \text{atan}(\tilde{\beta}_{t_0}^T \mathbf{f}) + \sum_{t=t_0+1}^T \text{atan}(\beta_t^T \mathbf{f}) \right\} \quad (3)$$

- Determine the *pseudo-label* of \mathbf{f}_i as $z_i = \text{sign}(H(\mathbf{f}_i))$
- Compute the weight for $i \in \mathcal{F}_u$: $w_i = e^{-z_i H(\mathbf{f}_i)}$, and normalize it to satisfy $\sum_{i \in \mathcal{F}_u} w_i = 1$.
- Compute the weighted error ϵ according to (4) and the gradient $\frac{d\epsilon}{d\tilde{\beta}_{t_0}}$ according to (5)
- If ϵ is decreasing then update $\tilde{\beta}_{t_0} \leftarrow \tilde{\beta}_{t_0} - \lambda \frac{d\epsilon}{d\tilde{\beta}_{t_0}}$ Else terminate the loop.
- **End**
- $\tilde{\beta}_{t_0} = \tilde{\beta}_{t_0}$

In the second step, we should utilize the unlabelled sample set \mathcal{F}_u to obtain an improved classifier. To this end, we need to define the so-called ‘‘pseudo-label’’ of unlabelled samples. There exist many approaches to define ‘‘pseudo-label’’, here we adopt a straightforward and efficient approach: Using existing classifier to define the ‘‘pseudo-label’’. Note that this is similar to the ‘‘self-learning’’ but it is in fact totally different with ‘‘self-learning’’. In ‘‘self-learning’’, the estimated ‘‘pseudo-label’’ is fixed during the whole training period but in our works, the ‘‘pseudo-label’’ changes in each iteration.

After then, the goal of the improvement is to minimize the weighted error

$$\epsilon = \sum_{i \in \mathcal{F}_u} w_i (H(\mathbf{f}_i) - z_i)^2 \quad (4)$$

where z_i is ‘‘pseudo-label’’ and w_i is the samples weight which will be defined later. Similar to [9], we adopt the gradient descent method to solve this problem iteratively.

Taking the derivative with respect to β_t gives

$$\frac{d\epsilon}{d\beta_t} = \frac{4}{\pi} \sum_{i \in \mathcal{F}_u} w_i (H(\mathbf{f}_i) - z_i) \frac{\mathbf{f}_i}{1 + (\beta_t^T \mathbf{f}_i)^2} \quad (5)$$

Then β_t can be updated as $\beta_t \leftarrow \beta_t - \lambda \frac{d\epsilon}{d\beta_t}$, where the step size λ should be determined by line search (this parameter was neglected by [9]). The updating process can be proceeded according to recursive form (see (3)). The update algorithm is summarized in Algorithm 2.

3 The Framework for Visual Tracking

We adopt the approach similar to [11] to design the online classifier, i.e., each pixel is regarded as a sample and the feature vector is constructed using its RGB value (therefore the feature vector can be formed as $[R \ G \ B \ 1]^T$).

3.1 Classifier Updating

During tracking period, we can never get labelled samples. Most of the “self-learning” approaches assume that “positive” samples can be extracted from the current tracking result and the “negative” samples can be extracted from the surrounding region. However, due to some unpredictable factors which often occur, such as occlusions, and temporary tracking failures. The labels of such samples may include noises, even are totally wrong. To tackle this problem, [11] proposed an approach for outlier rejection, which can avoid wrongly updating of classifier in some sense. However, this approach is based on a hard threshold, which is difficult to determine in practice. In this work, we regard these samples as unlabelled samples and use them to improve the existing classifier.

The concrete updating procedure follows Algorithm 2. Assume that at time instant $k - 1$ we have strong classifier

$$H_{k-1}(\mathbf{f}) = \frac{2}{\pi} \sum_{t=1}^T \text{atan}(\beta_{k-1,t}^T \mathbf{f}). \quad (6)$$

The tracking process is formulated into the framework of particle filter, which will be described later. At the new frame, we generate some particles and use H_{k-1} to evaluate these particles. Then the weighted sum of these particles is produced to obtain the current tracking result. After that, we use these particles which are randomly generated as the unlabelled samples and call Algorithm 2 to update the previous classifier to get the new classifier

$$H_k(\mathbf{f}) = \frac{2}{\pi} \sum_{t=1}^T \text{atan}(\beta_{k,t}^T \mathbf{f}). \quad (7)$$

3.2 Bayesian Tracking

The task of tracking is to use the available measurement information to estimate the hidden state variables. Given the available observations $\mathbf{z}_{1:k-1} = \mathbf{z}_1, \mathbf{z}_2, \dots, \mathbf{z}_{k-1}$ up to time instant $k - 1$, the prediction stage utilizes the probabilistic system transition model $p(\mathbf{x}_k | \mathbf{x}_{k-1})$ to predict the posterior at time instant k as $p(\mathbf{x}_k | \mathbf{z}_{1:k-1}) = \int p(\mathbf{x}_k | \mathbf{x}_{k-1}) p(\mathbf{x}_{k-1} | \mathbf{z}_{1:k-1}) d\mathbf{x}_{k-1}$. At time instant k , the observation \mathbf{z}_k is available, the state can be updated using *Bayes's* rule $p(\mathbf{x}_k | \mathbf{z}_{1:k}) = \frac{p(\mathbf{z}_k | \mathbf{x}_k) p(\mathbf{x}_k | \mathbf{z}_{1:k-1})}{p(\mathbf{z}_k | \mathbf{z}_{1:k-1})}$, where $p(\mathbf{z}_k | \mathbf{x}_k)$ is described by the observation equation. The kernel of particle filter is to recursively approximate the posterior distribution using a finite set of weighted samples. Each sample \mathbf{x}_k^i represents

one hypothetical state of the object, with a corresponding discrete sampling probability ω_k^i , which satisfies $\sum_{i=1}^N \omega_k^i = 1$. The posterior $p(\mathbf{x}_k | \mathbf{z}_{1:k})$ then can be approximated as $p(\mathbf{x}_k | \mathbf{z}_{1:k}) \approx \sum_{i=1}^N \omega_k^i \delta(\mathbf{x}_k - \mathbf{x}_k^i)$, where $\delta(\cdot)$ is Dirac function. Then the estimation of the state \mathbf{x}_k can be obtained as $\hat{\mathbf{x}}_k = \sum_{i=1}^N \omega_k^i \mathbf{x}_k^i$. The candidate samples $\{\mathbf{x}_k^i\}_{i=1,2,\dots,N}$ are drawn from an importance distribution $q(\mathbf{x}_k | \mathbf{x}_{1:k-1}, \mathbf{z}_{1:k})$ and the weight of the samples are $\omega_k^i = \omega_{k-1}^i \frac{p(\mathbf{z}_k | \mathbf{x}_k^i) p(\mathbf{x}_k^i | \mathbf{x}_{k-1}^i)}{q(\mathbf{x}_k | \mathbf{x}_{1:k-1}, \mathbf{z}_{1:k})}$. The samples are re-sampled to generate an unweighed particle set according to their importance weights to avoid degeneracy. In many cases, $q(\mathbf{x}_k | \mathbf{x}_{1:k-1}, \mathbf{z}_{1:k})$ is set to be $p(\mathbf{x}_k | \mathbf{x}_{k-1})$ and the weights therefore become proportional to the observation likelihood $p(\mathbf{z}_k | \mathbf{x}_k)$. In this paper, the observation likelihood is determined by the output of the online classifier:

$$p(\mathbf{z}_k | \mathbf{x}_k^i) \propto \frac{e^{H_{k-1}^i}}{e^{H_{k-1}^i} + e^{-H_{k-1}^i}} \quad (8)$$

where H_{k-1}^i is the output of the previous classifier on the current sample \mathbf{x}_k^i .

3.3 Comparison with Existing Approaches

In this section we give a detailed comparison between the proposed approach with the most related literature, i.e. [9], [12], and [6].

- The gradient feature selection approach was proposed by [9], where Histogram of Oriented Gradient (HOG) was used as the feature. Since the adopted HOG is multiple dimensional vector, [9] adopted weighted Linear Discriminative Analysis (LDA) to project it to 1-D feature space. However, in many cases, the number of samples is less than the number of dimensions and therefore LDA approach does not work. In this paper, we use weighted least square to get the project coefficient vector β_t and therefore avoid this problem. In addition, in the gradient learning of [9], the optimization variable includes not only the project coefficients, but also the coordinate values of the cells of which HOG is extracted. However, it can be noted that these variables are independent optimized according to gradient learning, without considering their mutual constraints. If these constraints are neglected, it is possible to get invalid solution. In this paper, what we optimize is just the project coefficient β_t and therefore no constraints need to be considered. Finally, the approach of [9] is not suitable for tracking small object since in this case HOG feature is difficult to get.
- In [12], an adaptive online boosting approach was proposed for tracking. However, [12] introduced many parameters to be determined by the designer. In addition, to avoid wrongly model updating, the online learning is switched on or off depending on a hard threshold, which is difficult to determine in practice. Finally, the approach proposed in [12] belongs to “self-learning” and our works belongs to “semi-supervised learning”.
- [6] used SemiBoost [11] to design a tracker. However, there exist three major differences between [6] and our work.

- The adopted features are different. In [6], Harr-like features are adopted. For small object, these features are difficult to extract. In this paper, the tracker works on pixel level.
- SemiBoost [11] was originally proposed for off-line application. It utilizes the similarity information between each pair of all samples. The calculating of similarity information is very time-consuming and therefore the online application of SemiBoost is restrictive. To tackle this problem, [6] avoided the similarity calculating by neglect the relationships between unlabelled samples and used a prior classifier to approximate the similarity between labelled samples and unlabelled samples. Therefore, the semi-supervised learning approach used in [6] is just an approximate version of SemiBoost. In fact, SemiBoost strongly emphasizes the importance of similarity information, which however plays little role in the work of [6]. In addition, a demerit of SemiBoost is that the parameter in similarity is difficult to determine. [6] claimed to propose an learning approach to determine the similarity. However, there still exist parameters to be determined in calculating similarity. In [6], the classifier is updated once one unlabelled sample is extracted and therefore the relationships between unlabelled samples are totally neglected. In fact, in tracking scenarios, we can get many unlabelled samples in batch manner for ONE frame and can sufficiently utilize the distribution information to update the classifier.
- Similar to [5], [6] randomly selected the features and does not fully take advantage of the nature that for object tracking, i.e., the high correlation over time. While in this paper, the gradient learning approach which exploits the correlation of sequential data is used for online learning.

4 Experimental Results

In this section, we evaluate the proposed tracker on some video sequences showing that the semi-supervised approach can improve tracking performance. In our algorithm, the number of weak classifiers is set to $T = 10$. During tracking period, 300 pixels around the current tracking results are randomly extracted in each frame to be unlabelled samples.

The first examples is tested on color video sequence from OTCBVS dataset collection (<http://www.cse.ohio-state.edu/OTCBVS-BENCH>) [4]. In this example,

For the same video sequence, we make a comparison between the tracking approach of [11] and ours. In Fig.1, the first column gives the initialization results. The second and third column correspond to two adjacent frames. It shows that both the approaches in [11] and ours give deflected tracking results at frame 51. Since [11] adopts a “self-learning” approach, the classifier will then be wrongly updated and the performance of the classifier deteriorates from then on (see the third and fourth columns). On the other hand, since we use online semi-supervised technology, we donot use the hard label of the extracted samples. Therefore the error will no accumulate and the performance recovers during the next frames.

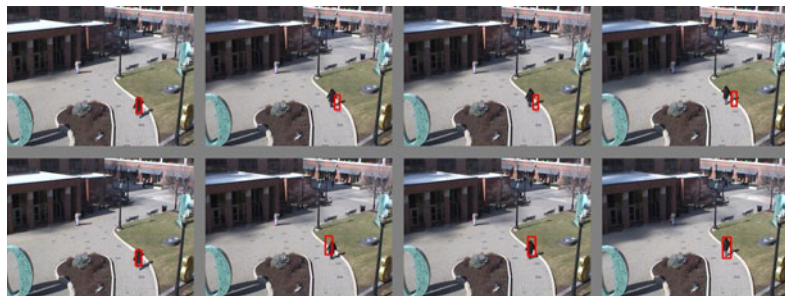


Fig. 1. From left to right: Frames 2, 51, 52 and 72. The first row corresponds to ensemble tracking [1]; The second row corresponds to the approach proposed in this paper.

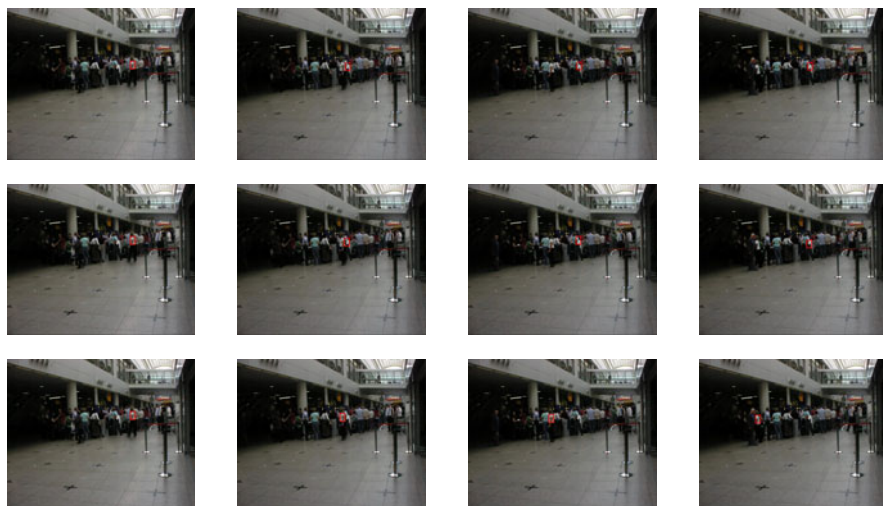


Fig. 2. From left to right: Frames 2, 43, 70 and 98. The first row corresponds to ensemble tracking [1]; The second row corresponds to the supervised version of the proposed approach (i.e. **GSL**). The third row corresponds to the approach proposed in this paper.

We also use PETS2007 dataset for tracking demonstrations. In Fig.2 we show some representative images. In this experiment, we attempt to track the head of a woman through clutters. In addition to ensemble tracking approach [1], we also develop a supervised version of the proposed approach. This supervised approach is similar to Algorithm 2, except that the labels of samples are known and fixed — The labels are extracted according to self-learning approach similar to [1]. For convenience, we call this as “Gradient Supervised Learning” (**GSL**) and Algorithm 2 as “Gradient Semi-Supervised Learning” (**GS-SL**). By this

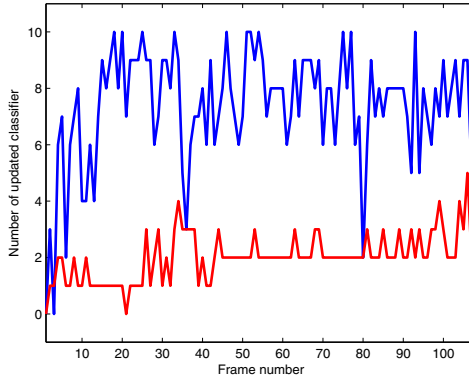


Fig. 3. The number of weak classifiers that change their parameters over time. (Red line: **GS-SL**; Blue line: **GSL**)

comparison we can show the advantages of **GS-SL**. All three approaches begin with the same initialization results. At frame 43, the result of ensemble tracking (the first row) is distracted by a nearby white object, while **GS-SL** (the third row) is not influenced by it. After frame 43, the performance of ensemble tracker never recovers and **GS-SL** succeeds in tracking till frame 98. Also, **GSL** (the second row) achieves similar results to ensemble tracking.

Finally, we give the number of updated classifiers of **GSL** and **GS-SL** over time in Fig.3 for comparison. From this figure we can see that in most cases **GS-SL** updates less classifiers than **GSL**. This illustrates that semi-supervised learning provides a more “mild” update strategy.

5 Conclusions

In this paper, we developed a gradient semi-supervised learning approaches for two applications. First, at the first frame, usually we can get an initialization result which is given by an extra detector, or by manual labelling. However, these results usually include wrongly labelled samples. So we adopt semi-supervised learning technology to refine the initial classifier. Secondly, during the tracking period, the semi-supervised technology is used to online update the classifier. In addition, the classifier update procedure is based on gradient learning which exploits the correlation of sequential data.

Acknowledgements

This work is jointly supported by National Key Project for Basic Research of China (Grant No. G2007CB311003), Natural Science Foundation of China (Grants No. 61075027, 90820304), and Natural Science Foundation of Hebei Province (Grant No. F2010001106).

References

1. Avidan, S.: Ensemble tracking. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 261–271 (2007)
2. Chen, H., Liu, T., Fuh, C.: Probabilistic tracking with adaptive feature selection. In: *Proc. of Int. Conf. on Pattern Recognition*, pp. 736–739 (2004)
3. Collins, R., Liu, Y., Leordeanu, M.: Online selection of discriminative tracking features. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 1631–1643 (2005)
4. Davis, J., Sharma, V.: Fusion-based background subtraction using contour saliency. In: *Proc. of Int. Workshop on Object Tracking and Classification Beyond the Visible Spectrum*, pp. 1–8 (2005)
5. Grabner, H., Bischof, H.: On-line boosting and vision. In: *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pp. 260–267 (2006)
6. Grabner, H., Leistner, C., Bischof, H.: Semi-supervised on-line boosting for robust tracking. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008, Part I. LNCS*, vol. 5302, pp. 234–247. Springer, Heidelberg (2008)
7. Hu, W., Tan, T., Wang, L., Maybank, S.: A survey on visual surveillance of object motion and behaviors. *IEEE Trans. on Systems, Man and Cybernetics*, 334–352 (2004)
8. Li, Y., Ai, H., Yamashita, T., Lao, S., Kawade, M.: Tracking in low frame rate video: A cascade particle filter with discriminative observers of different lifespans. In: *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pp. 1–8 (2007)
9. Liu, X., Yu, T.: Gradient feature selection for online boosting. In: *Proc. of Int. Conf. on Computer Vision*, pp. 1–8 (2007)
10. Lucey, S.: Enforcing non-positive weights for stable support vector tracking. In: *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pp. 1–8 (2008)
11. Mallapragada, P., Jin, R., Jain, A., Liu, Y.: Semiboost: Boosting for semisupervised learning. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 2000–2014 (2009)
12. Parag, T., Porikli, F., Elgammal, A.: Boosting adaptive linear weak classifiers for online learning and tracking. In: *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pp. 1–8 (2008)
13. Tang, F., Brennan, S., Zhao, Q., Tao, H.: Co-tracking using semi-supervised support vector machines. In: *Proc. of Int. Conf. on Computer Vision*, pp. 1–8 (2007)
14. Wang, J., Chen, X., Gao, W.: Online selecting discriminative tracking features using particle filter. In: *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pp. 1037–1042 (2005)
15. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. *ACM Computing Surveys*, 1–45 (2006)
16. Zhang, C., Hamid, R., Zhang, Z.: Taylor expansion based classifier adaptation: Application to person detection. In: *Proc. of Int. Conf. on Computer Vision and Pattern Recognition*, pp. 1–8 (2008)
17. Zhu, X.: Semi-supervised learning literature survey, Technical Report, University of Wisconsin-Madison (2005)

Solving Multiple-Target Tracking Using Adaptive Filters

B. Cancela, M. Ortega, Manuel G. Penedo, and A. Fernández

Varpa Group, Department of Computer Science

University of A Coruña, Spain

{brais.cancela,mortega,mgpenedo,alba.fernandez}@udc.es

Abstract. Multiple-target tracking represents a challenging question in uncontrolled scenarios. Due to high-level applications, such as behavioral analysis, the need of a robust tracking system is high. In a multiple tracking scenario it is necessary to consider and resolve occlusions, as well as formations and splitting of object groups. In this work, a method based in a hierarchical architecture for multiple tracking is proposed to deal with these matters. Background subtraction, blob detection, low-level tracking, collision detection and high-level appearance tracking is used to avoid occlusion and grouping problems. Experimental results show promising results in tracking management, grouping, splitting, occlusion events, while remains invariant to illumination changes.

1 Introduction

Multiple target tracking around a scene has become one of the most active research fields nowadays. This is so because of the need of techniques to capture object behavior in applications such as video surveillance or body capturing. Despite this interest, it is a problem far to be solved, specially under uncontrolled scenarios. Object tracking tries to recognize and list non-rigid objects under illumination and background surface changes. Due to the type of use, every object is considered as a solid blob, thus tracking objects surrounded by other is forbidden. Tracking objects also interact with others, and it is interesting to define such relations in order to introduce high level reasoning over object behavior. It is also necessary because of grouping events. When an object collides with another one, it is difficult to track both objects unless you can detect the collision introducing object groups to model the inner collection. Once the objects are separated in the scene, the split objects should be recognized as the old ones in the scene prior to the collision.

In this work we present a methodology for multiple-object tracking that deals with those situations, establishing an adequate framework for higher-level applications that require tracking, such as pedestrian trajectory analysis. This paper is organized as follows: section 2 shows different approaches to the problem, including our method; section 3 describes low level tracking, whereas section 4 describe high-level tracker and target appearance detection; section 5 shows some experimental results and section 6 offers conclusions and future work.

2 Related Work

There are in the literature different approaches for object tracking. Low-level approaches obtained good results tracking isolated objects. In [1] an optical flow detection is used to track vehicles in an automatic video surveillance system, obtaining bad values over non-moving objects. Rohr [2] uses Kalman Filters to predict target position under noise conditions. However, they cannot solve the occlusion problem.

High-level approaches try to learn complex templates a priori in order to do pattern matching. BraMBLe [3] models both background and foreground using Mixture of Gaussians. Particle Filters are also a common choice ([4], [5]), but it cannot deal with collision events and no multiple-tracking events are considered. Brand et al. [6] use a coupled Hidden Markov Models to determine object interactions. M. Li et al. [7] introduced a tracking based in omega-shape features, which performs a multiple tracking-people by head-shoulder pattern detection.

Our approach to this topic is based in the hierarchical architecture proposed by Daniel Rowe et al. [8]. In that work, each level performs a distinct functionality. The lower level performs target detection, consists in background subtraction and a blob detection using foreground contours. For each detection, the following level obtains an ellipse representation for each object. Later, they reduce its appearance using color histograms, which are less sensitive to rotations in depth or target deformations. A bunch of 49 different RGB linear combinations is reduced to M histograms, according to a foreground/background ratio. Finally, last level establishes coherent target relations between frames, including grouping, splitting and leaving the scene cases.

Our goal is to improve this hierarchical architecture to deal in a better way with the problems mentioned before. As a result, a robust system is presented, with background independent position and appearance information, with a better approach to the illumination invariance. Qualitative information, such as object relations or occlusions, will also be presented. Our methodology works as follows: first, a background subtraction is computed. Second, interest regions are detected. Third, an object representation is obtained. Finally, coherent target relations are established between frames.

3 Blob Detection and Low-Level Tracking

The first stage of our system is the detection of blobs within the scene. In [8], a method proposed by Horprasert et al. [9], based on a color background-subtraction, is used. However, this method requires an extra parameter, $\tau_{\alpha lo}$, to locate dark foreground values. Since there is no automatic method to calculate $\tau_{\alpha lo}$, this approach is very dependent of lighting condition. Thus, a lot of noise is introduced when it is not well calibrated.

To avoid this problem, we use a method based in Mixture of Gaussians [10] (MoG). Each background pixel is modeled as a bunch of gaussians, typically three. Pixel values that do not fit the background distributions are considered foreground until there is a Gaussian that includes them with enough evidence. Background model is updated with every new frame using a parameter, α , which is the learning rate. Often every stopped blob may become part of the background after some updates. To prevent this

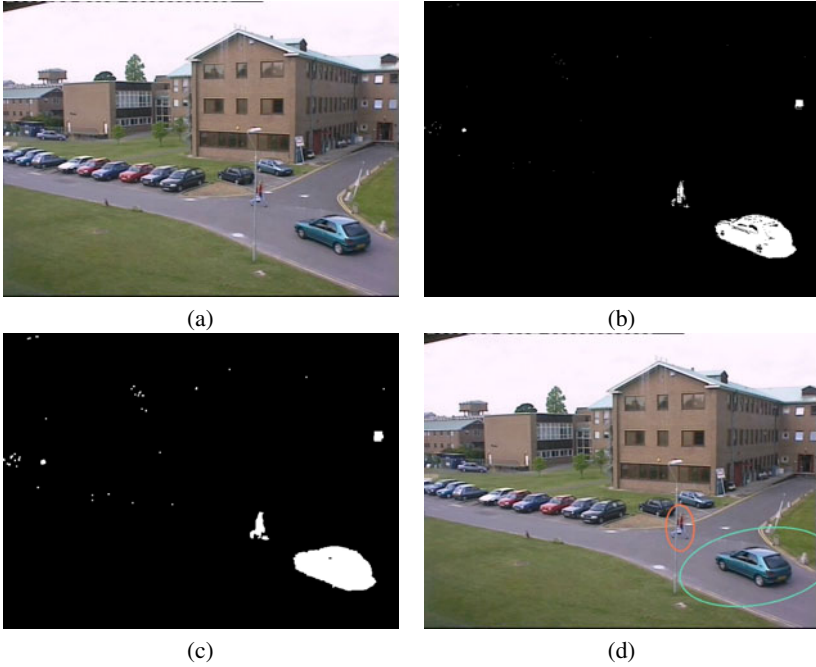


Fig. 1. (a) Frame. (b) MoG foreground detection. (c) Morphological operators. (d) Blob detection after applying minimum-area filter.

situation, the training parameter α is set to a very low value. A requirement is that an only-background sequence is needed to train the algorithm.

Once our background subtraction methodology is run, we have pixels classified into two categories: foreground and background. After this, we apply dilation and closing operators with a minimum-area filter in order to fill the blobs and avoid small regions, holes or noise due to alterations such as camera movement or video compression when detecting blob regions. Once detected, the j -observed blob at time t is given by $z_j^t = (x_j^t, y_j^t, h_j^t, w_j^t, \theta_j^t)$, where x_j^t, y_j^t represent the ellipse centroid, h_j^t, w_j^t are the major and minor axes, and θ_j^t the ellipse orientation. Fig. 1 shows an example of this methodology.

To prevent from noisy measures, a bunch of Kalman Filters is used to predict the target state. Then, measure validation is established according the regions where the target observations are expected. A specific Mahalanobis Square Distance (MSD), using the innovation covariance matrix S_k , is computed to set the gates. This method obtains good results, however, in absence of noise the MSD value tends to zero, even with far ellipses from the predicted position. Furthermore, when a tracked object is lost, predicted state tends to diverge in few iterations. To solve this situation, our approach stores the position of the ellipse in a window of size M . Subsequently, a median filter is used in order to smooth the values and a set of adaptive filters (Adalines) predict the velocity of each ellipse parameter. Once the velocity is computed, it is added to the previous position to obtain the prediction. With a low training parameter, this method guarantees a smooth prediction in case a tracking object is lost.

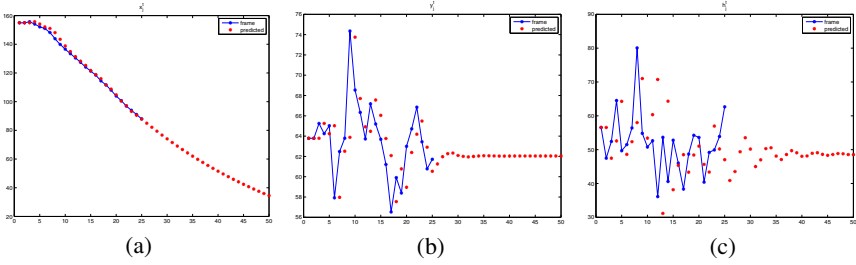


Fig. 2. (a) x_j^t component. (b) y_j^t component. (c) h_j^t component. Blue color represents frame values and red predictions. System is robust under occlusion events.

Fig. 2 shows filter response for each component of a certain object. x_j^t is easily predicted and the adaptive filter obtains good results. When the object becomes occluded, it also shows promising values. Furthermore, under noise conditions, such as y_j^t and h_j^t components, stable values are guaranteed.

The method to match the ellipse frame with an existing tracking object starts by computing the predicted state of the tracking object. We assumed that every object in the frame moves slowly enough compared to the frame rate. Hence, if the ellipse frame centroid is located within the predicted state, the low level tracking is confirmed.

To locate the position of the centroid i, j with respect to the ellipse $z_j^t = (x_j^t, y_j^t, h_j^t, w_j^t, \theta_j^t)$ a transform is applied as follows:

$$\begin{bmatrix} i_z \\ j_z \\ 1 \end{bmatrix} = \begin{bmatrix} \cos\theta_j^t & -\sin\theta_j^t & 0 \\ \sin\theta_j^t & \cos\theta_j^t & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & -x_j^t \\ 0 & 1 & -y_j^t \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} i \\ j \\ 1 \end{bmatrix}, \tag{1}$$

where (i_z, j_z) represents (i, j) under z_j^t coordinates. If $\frac{i_z^2}{h_j^t} + \frac{j_z^2}{w_j^t} \leq 1$, the centroid is within the ellipse.

4 High-Level Appearance Tracker

As mentioned before, low-level tracking has problems in cases of grouping and occlusion, since it cannot detect when two tracking objects become a group, or vice versa. To solve this, Rowe et al. propose the implementation of high-level trackers which include information relative to the tracking appearance. Due to light sources, orientation and position changes in the tracking object, the appearance must be updated every iteration.

In [11], an appearance modeling approach is presented. This method uses multiple color features, which are evaluated and ranked taking into account the background near the object. With multiple combinations between the RGB components, they obtain 49 different lineal dependent histograms. After this, the best N histograms are selected, according to the differences between foreground histogram and local background histogram. The selected N histograms should be similar between frames. However, this is not true under noise conditions or after occlusions, because the background difference between the last and new detection could be high. Thus, based on this initial approach,

we propose a definition of a fixed pool of valid histograms even if the background or the illumination change.

Our proposal for high-level tracking consists of four main steps: management of states in order to model different interactions of objects (collision, splittings, occlusions), tracking object matching, feature selection and appearance computation.

4.1 Object State Management

Six different states are defined: single target, target grouping, grouped, splitting, split and occluded. Once the ellipses in the new frame are computed and the predicted positions of the tracked target are calculated, a collision detection procedure is computed.

First, if two or more different ellipse centroids in the frame are within the same tracking object predicted position, the target change its state to *splitting*. If the split is confirmed in the next frame the state is changed to *split*. Then, the tracking object is removed. If it is a group, we compare the appearance between the objects involved in the group and the new ellipses. When a match is obtained, we associate the new ellipse to the matching object. Finally, if we have any blobs left and we have no unmatched tracking object associated with the group, we instantiate a new tracking object for every ellipse within the group. In the other hand, if there are any tracking object associated with the group and no blobs, we change every object state into *occluded*. If, on the contrary, there are both blobs left and unassociated tracking objects, we instantiate a new *group* for every blob we have, and every tracking object is associated to every group. If the split is not confirmed, the tracking object state is changed to its previous value.

Second, if two or more different tracking object are predicted within a ellipse in the frame, a new tracking object is created with its state moved to *grouping*. If the group is confirmed, the state is changed to *grouped*. Then, every tracking object within the ellipse in the frame as part of the group is associated to the new tracking object. None of the them can be used out of the group unless the group is dissolved.

Finally, if a tracking object labeled as *grouped* does not match the appearance with a blob or a tracking object associated to the group has a more similar appearance, we remove the group and label every tracking object as *occluded* except the matched one.

4.2 Tracking Object Matching

This module is activated when a low-level tracking object is trained. This happens when it is detected for six consecutive frames. In previous approaches tracking object appearance is calculated whenever the low-level tracker is confirmed. In our case, we compute tracking appearance in all cases, because we can train features more quickly. In the case of a new tracking object, in the first steps, both high-level and low-level parameters relative to the target position and shape are updated. However, until we the low-level tracking is not considered trained, we do not compute a feature comparison, because first we need to obtain a good appearance representation, avoiding noise problems. Furthermore, when the tracking object is trained, appearance comparison will also computed.

The system associates every tracking object to the newly created ones. If the high-level appearance tracker confirms the match, the state is updated with the new position. If no observation is associated to a particular target, its state is set using the previously predicted state.

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

4.3 Feature Selection

The tracking object space is represented using color histograms. Instead of using raw R, G, and B channels, we propose to use L*a*b, which is a color-opponent space with dimension L for lightness and a and b for the color-opponent dimensions. This way we can isolate the illumination into one component in order to work with the other two. Therefore, we compute the histograms as follows

$$h = \omega_1 * a + \omega_2 * b, \quad \omega_{1,2} \in \{-1, 0, 1\} \quad (2)$$

If we avoid possible lineal combinations between them, we reduce 8 to 4 different histograms. In some cases, the lightness is important in order to make a distinction between two different objects, so we add one more histogram calculated with the L component. Features are normalized and discretized into 64 bins, which is high enough to prevent from wrong matching, according to [8]. Thus, the i th-feature tracking object histogram is given by $\mathbf{p}^i = \{p_k^i; k = 1 : 64\}$. The probability of each feature is calculated as:

$$p_k^i = C_i \sum_{a=1}^M \delta(b(x_a) - k), \quad (3)$$

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

4.4 Appearance Computation

For each one of the five features, the mean appearance histogram of the i th-feature in time t , \mathbf{m}_t^i , is recursively computed:

$$\mathbf{m}_t^i = \frac{n_i \mathbf{m}_{t-1}^i + \mathbf{p}_t^i}{n_i + 1}, \quad (4)$$

where n_i is the number of times the histogram has been computed. Similarity between two histograms is computed using the *Hellinger distance*, $d_H = \sum_{k=1}^{64} \sqrt{p_k q_k}$. Therefore, the mean and variance of d_H are updated with every new match. In fact, features of a new frame target are matched using the previous mean and variance. If the frame target matches with an existing tracker, this is updated. We consider a match to occur when at least the 60% of the comparisons pass the test. If there is no match, a new

tracking object is instantiated and trained. Once this is trained, it is compared against other *occluded* trackers, because it could be one of the previously defined. If this is the case, the tracker is merged with the previous one.

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

5 Experimental Results

In our experiments we have used the PETS 2001 Test Case Scenarios [12] in order to test the methodology. Three different videos are used, which implies more than four minutes video recording at 25 frames per second. These videos take place outdoors. Partial occlusions, grouping and splitting events are evaluated, as well as target exiting and entering into the scene.

The sequences used in this algorithm test involve isolated people, groups and vehicles. Once isolated targets are detected and each tracker is trained, the system performs well both grouping and splitting events between them. In Fig. 3 we can see a crossing example. Two trackers are trained and instantiated. At some point, the path of both targets is crossed, which implies the creation of a new group tracker including both of them. Later, the trackers are far enough and the split is performed. The algorithm detects the new positions and assigns in a correct way the identification of the two objects. There are cases in which a group stayed close to the end of camera range. If one member of the group leaves the scene, the algorithm detects that the new appearance is close to one of the members rather than the group and remove it. One example of this problem can be watched in Fig. 4. However, the method cannot handle the case of one member close to the end of camera range and other object appeared just before it. Once the target are far enough, the split method is activated and the algorithm process the first target as a group.

The case of occluded target is also referred. As we see in Fig. 5 the system can recover target identification under short occlusions, around 100-150 frames. The algorithm performs well under total and partial occlusions. In cases of objects leaving the scene and reentering much later (more than 150 frames), it obtains 50% of recovery rate.

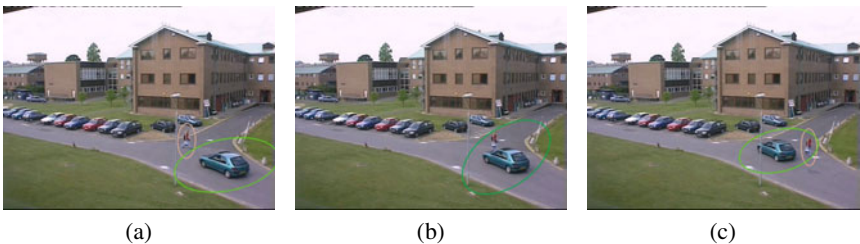


Fig. 3. (a) Tracking before grouping. (b) Tracking during grouping. (c) Tracking after grouping. Ellipse color represents tracking identification. Once the group is created, a new tracker is instantiated containing the two trackers. When splitting occurs, both trackers have the same previous identification.

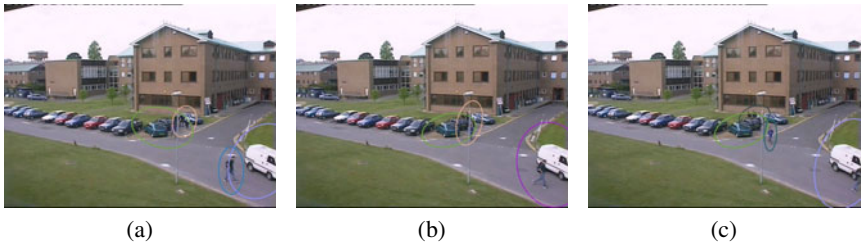


Fig. 4. (a) Tracking before grouping. (b) Tracking during grouping. (c) Tracking after human leaves the scene. The new ellipse appearance computation shows that the new object is closer to the car than to the group, so the group is removed and the object is associated to the car identification.



Fig. 5. Tracking object with occlusions. The algorithm can recover tracker identification over occlusions with less than five seconds (100-150 frames).

Table I shows the results, mainly focused into object interactions. The object detection performs well when using MoG algorithm, which implies that, once tracker is trained, grouping events always have good results. Splitting events are generally well detected but some problems arise in cases of more than two targets in a group. This is because when there are more than two humans walking next to each other, split algorithm have problems to make stable trackers to each person individually. In general, short time occlusion events are good performed, meanwhile long time sometimes have problems derived to appearance model.

Table 1. Multiple-Target tracking results. Grouping and splitting method performs really well. Most of the errors assumed is derived by background subtraction and blob detection. Illumination changes and leaf movements introduce bad tracking objects. The system is able to recover a target in more than a half of occlusion cases. Lost objects are tracking interrupt due to appearance bad matching. MoG also introduces noise in the background subtraction, and the minimum-area filter occasionally lost targets, which are moving away from the camera.

	Total	Correct	Incorrect	%
Tracked objects detected	24	18	6	75
Grouping events	18	18	0	100
Splitting events	24	21	3	87.5
Occlusion recovery	12	7	5	58.3

6 Conclusions

In this paper a new approach to the hierarchical architecture is presented in order to avoid some of the problems presented in this architecture. A different low-level tracking, based in adaptive filters and ellipse formulation, is implemented with good results. An appearance model based in L^*a^*b color space is preformed to obtain a background independent appearance target model, showing promising results. Robust tracking is achieved, even under grouping and splitting situations. In cases of sudden illumination changes, background subtraction algorithm fails and blob detection is not feasible, so constant illumination is needed in order to obtain good results.

In a future research a different blob detection would be interesting in order to obtain a fully illumination independent algorithm. Also a new method to achieve a good recovery identification under long occlusions is desired. A methodology for behavior analysis in a future work, such as trajectory analysis, will be developed.

Acknowledgments

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

References

- [1] Huang, T., Koller, D., Malik, J., Ogasawara, G.H., Rao, B., Russell, S.J., Weber, J.: Automatic symbolic traffic scene analysis using belief networks. In: Proc. Nat. Conf. Artif. Intell., pp. 966–972 (1994)
- [2] Rohr, K.: Towards model-based recognition of human movements in image sequences. CVGIP, Image Understanding 59(1), 94–115 (1994)
- [3] Isard, M., MacCormick, J.: Bramble: A bayesian multiple-blob tracker. In: 8th IEEE ICCV, Vancouver, Canada, vol. 2, pp. 34–41 (2001)
- [4] Kitagawa, G.: Monte carlo filter and smoother for non-gaussian nonlinear state space models. Journal of Computational and Graphical Statistics 5(1), 1–25 (1996)

- [5] Nummiaroa, K., Koller-Meierb, E., Van Gool, L.: An adaptive color-based particle filter. *Image and Vision Computing* 21(1), 99–110 (2003)
- [6] Brand, M., Oliver, N., Pentland, A.: Coupled hidden markov models for complex action recognition. In: *Proc. IEEE Conf. Comput. Vis. Pattern Recognition*, pp. 994–999 (1997)
- [7] Li, M., Zhang, Z., Huang, K., Tan, T.: Rapid and robust human detection and tracking based on omega-shape features. In: *16th IEEE International Conference on Image Processing (ICIP)*, pp. 2545–2548 (2009)
- [8] Rowe, D., Reid, I., González, J., Villanueva, J.J.: Unconstrained multiple-people tracking. In: Franke, K., Müller, K.-R., Nickolay, B., Schäfer, R. (eds.) *DAGM 2006. LNCS*, vol. 4174, pp. 505–514. Springer, Heidelberg (2006)
- [9] Horprasert, T., Hardwood, D., Davis, L.S.: A robust background subtraction and shadow detection. In: *4th ACCV, Taipei, Taiwan*, vol. 1, pp. 34–41 (2000)
- [10] Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2, pp. 246–252 (1999)
- [11] Collins, R., Liu, Y., Leordeanu, M.: Online selection of discriminative tracking features. *PAMI* 27(10), 1631–1643 (2005)
- [12] PETS, *International Workshops Performance on Evaluation of Tracking and Surveillance* (2001), <http://peipa.essex.ac.uk/ipa/pix/pets/>

From Optical Flow to Tracking Objects on Movie Videos

Nhat-Tan Nguyen, Alexandra Branzan-Albu, and Denis Laurendeau

Computer Vision and Systems Laboratory,
Laval University, Quebec (Quebec) G1K 7P4 Canada

ntnguyen@gel.ulaval.ca

aalbu@ece.uvic.ca

denis.laurendeau@gel.ulaval.ca

<http://vision.gel.ulaval.ca/>

Abstract. This paper addresses the problem of tracking human motion in a movie sequence involving camera movement. We have developed an approach to track the bounding box of a human in motion without using any particular model. This method exploits motion vector fields from the image, then subtracts the motion caused by the camera to obtain the segmentation of the object. In addition, we introduce a multi-level tracking approach. This approach makes the tracking operation more robust, and less prone to errors. Experiments with movie sequences representing human walk are reported.

Keywords: optical flow, object tracking, motion vector fields, movie sequences, image processing.

1 Introduction and Related Work

Various methods have been proposed to address the problem of tracking human motion. We can classify these approaches into two categories. The first category uses 2D or 3D models such as skeletons with connected segments, volumetric human models like spheres or cylinders [1], [2]. It provides an effective representation of the physical structure and constraints of the human body, but leads to complex analytical computations by fitting and matching the articulated model. The second category uses spatio-temporal information. The spatio-temporal XT-slices are exploited from the video sequence volume XYT, then typical trajectory patterns can be associated with articulated motion [3]. Polana et al. [4] use the motion fields computed between successive frames to segment and track actors. A particle filter is proposed in [5] for visual tracking.

In this paper, we present a tool based on low-level information to detect and track objects in movies, especially when camera movement is involved. This tool is useful for subtitling. The remainder of the paper is organized as follows. Section 2 presents the process of color coding for motion vector fields. The background subtraction method is presented in Section 3. Section 4 illustrates how the algorithm can be used to detect and track objects in movies. Section 5 concludes the paper and outlines areas for future work.

2 Color Coding for Motion Vector Fields

A motion vector fields of a frame is obtained by calculating the optical flow between two successive frames. Each pixel in the fields contains a motion vector. In order to represent a motion vector fields in color, a color wheel is generated as illustrated in Fig.1. This color wheel is based on the HSV color representation. It is first formed by 3 basic colors: red, blue and green. Then, it is combined with the 3 primary colors: cyan, magenta and yellow. Basic colors and primary colors are paired in the following way: red and cyan, green and magenta, blue and yellow. The distances (in hue) between these colors are equal.



Fig. 1. The color system used for motion vector fields coding

Subsequently, the color is used to code the direction of the motion vector. For example, blue indicates 0 degrees, red 120 degrees and so on. The radius reveals the magnitude of the vector, and of course, the magnitude of the motion vector fields is normalized into the $[0,1]$ range. With a color wheel as in Fig.1, a colormap is constructed. A colormap C is a 360-by-3 matrix of real numbers between 0.0 and 1.0. Each row is an RGB vector that defines one color which corresponds to one angular degree in the color wheel. Given a motion vector $\mathbf{a} = (u, v)$ at pixel i , its direction and magnitude are defined respectively as $d = \arctan(u, v)/\pi$ and $m = \sqrt{u^2 + v^2}$, $m \in [0, 1]$. The mapping vector \mathbf{a} into the colormap is given by:

$$ind = \left\lfloor \frac{d + 1}{2}(360 - 1) + 1 \right\rfloor \tag{1}$$

where ind is an index of the colormap. Then the color representing the vector \mathbf{a} is $m \times C(ind)$. The magnitude m serves as the factor of intensity.

With this mapping, a motion vector fields can be displayed as a color image. Fig.2 presents some samples extracted from a movie sequence and its motion vector fields. The first row represents the case where the camera is panning to the right, the second row represents the case in which the camera is panning to

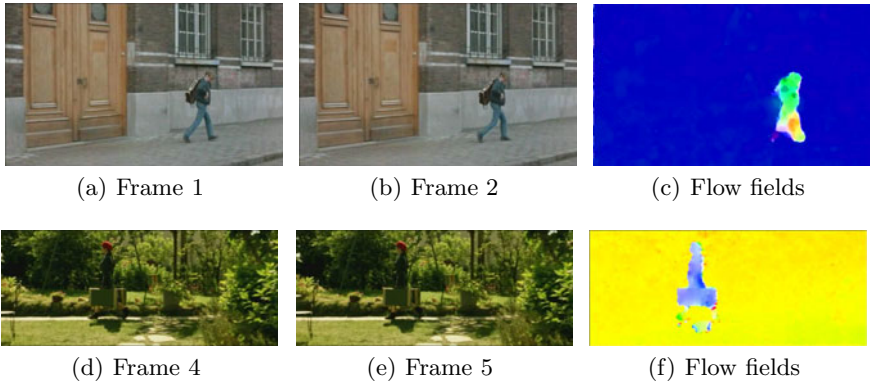


Fig. 2. Some samples and resulting motion vector fields

the left. For example, Fig. 2(c) is the colored flow fields image of the two successive images in Fig. 2(a) and 2(b). A pixel's value in the image in Fig. 2(c) contains the movement information of a corresponding pixel in the image in Fig. 2(b). In this video sequence, the camera is panning from left to right. Consequently, background pixels are all blue corresponding to the color wheel in Fig. 1. Fig. 2(f) is the colored flow fields image of the two successive images (in Fig. 2(d) and 2(e)) extracted from another sequence. This time, the camera is panning from the right to left and, therefore, background pixels are coded in yellow according to the color wheel.

3 Background Subtraction

The optical flow method is exploited based on the approach proposed by Zach et al. [6]. When the camera is panning or tilting, the object is usually distinguished from the background in the motion vector fields as shown in Fig. 2. In other words, the motion vectors that belong to the background are in the same direction, and, obviously, appear in the same color in the colored motion vector field image. In order to simplify the background subtraction step, blue is chosen as the coding color to indicate the moving background regardless of the direction of camera motion. To do this, the color wheel is rotated around its center through an angle in which the blue in the wheel is parallel to the direction of camera motion. This direction is known based on the camera motion estimation process described in [7]. Then the motion vector fields is coded with the rotated color wheel to obtain the colored motion vector field images. For example, the motion vector fields shown in Fig. 2(c) and Fig. 2(f) are re-coded with their own rotated color wheels are presented in Fig. 3. Fig. 3(a) and 3(b) represents the case where the camera is panning to the right and Fig. 3(c) and 3(d) represents the case in which the camera is panning to the left. The arrows on color wheels indicate the direction of camera motion.

With this kind of color image, a simple technique is used to eliminate the background. First, the color image of the motion vector fields is separated into

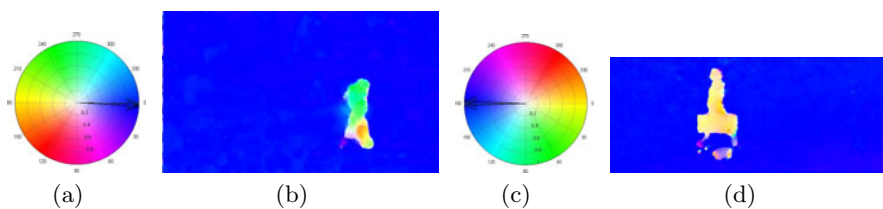


Fig. 3. Rotated color wheel and colored motion vector fields

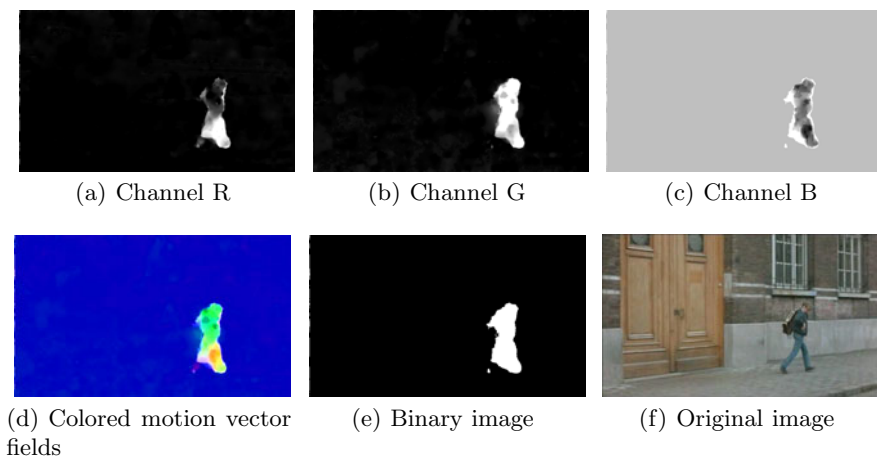


Fig. 4. Background subtraction results

three channels R , G and B (see Fig. 4(a), 4(b) and 4(c)). In each channel image (a grayscale image), Otsu's method [8] is applied to select the threshold that minimizes the intra-class variance of the black and white pixels and convert this grayscale image into a binary image. Given that R , G and B are binary images of the red, green and blue channels, then combining these images as $R \vee G \vee B$, a binary image I is obtained as presented in Fig. 4(e).

4 Multi-level Tracking

Adapted from the work of Torresan et al. in [9], the tracking is performed at multiple levels. The multi-level tracking algorithm is based on the concepts shown in Fig. 5(a), Fig. 5(b) which summarizes the overall tracking strategy.

At the first level of the procedure, the algorithm matches and groups one or several segmented regions, which are found at the background subtraction stage, to create *blobs*. This *blob tracking* level is processed in chronological order, i.e. from time t to $t + 1$, $t + 2, \dots$, named *Forward blob tracking*; and in reverse chronological order, i.e. from time t back to $t - 1$, $t - 2, \dots$, named *Backward blob tracking*. This strategy can be used since this project is not submitted to real-time tracking constraints. The results from these two processes are merged together (see Figs. 6(c) and 6(d)). The second level of tracking uses the results of

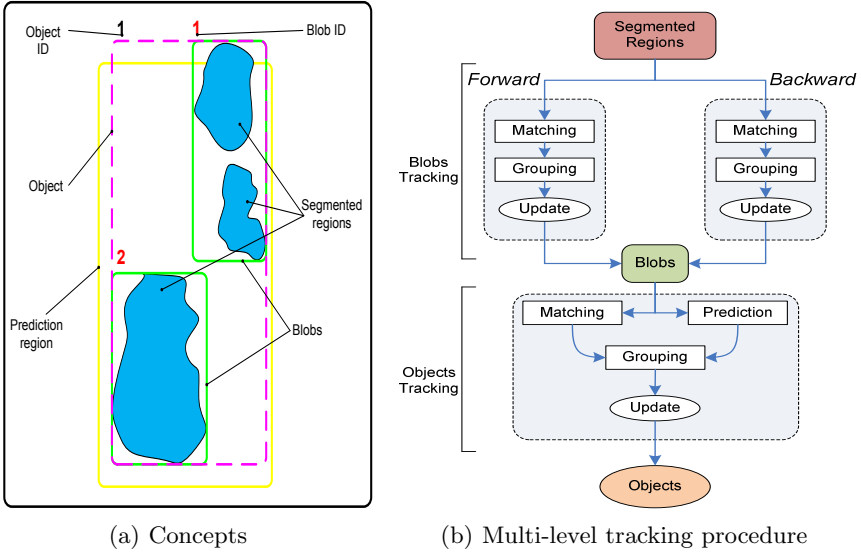


Fig. 5. Multi-level concepts and tracking strategy

this merging. The blobs detected at the first level are then matched with blobs in the next frame, compared to the prediction regions and grouped together to obtain the final labeled *objects*. The rest of this section will present details which are relevant to the tracking procedure.

Let a and b denote a blob at the time t and $t - 1$ respectively. Then $R(a, b)$ denoting the overlap between blob a and b , is defined formally as:

$$R_{max}(a, b) = Max \left(\frac{SC(a, b)}{RI(a)}, \frac{SC(a, b)}{RI(b)} \right) \tag{2}$$

$$R_{min}(a, b) = Min \left(\frac{SC(a, b)}{RI(a)}, \frac{SC(a, b)}{RI(b)} \right) \tag{3}$$

where $RI(i)$ is the area of the i^{th} blob's ROI and $SC(a, b)$ is the intersection area between the two ROI.

$S(a, b)$ is the similarity of blobs a and b :

$$S(a, b) = 1 - \left[\frac{Abs(SR_a - SR_b)}{Max(SR_a, SR_b)} \right] \tag{4}$$

where SR_i is the actual area of the i^{th} blob as indicated in Fig. 5(a).

During the tracking process, the maximum overlapping factor R_{max} is used to follow-up blobs between two successive frames of a sequence. In the meanwhile, R_{min} and S are used to reduce the correspondence between blobs.

At the second tracking level (i.e. *object tracking* level), one or more blobs can be grouped to create an *object*. The object's velocity is computed and is

then employed to produce a prediction region of the object. The position of the prediction region can be modified by the average speed of the prediction region of the last fifteen frames. Using the same principle, the dimension of the prediction region is obtained by computing the mean value of the prediction region of the last fifteen frames. When a new frame is processed, a blob that appears in a prediction region of an object can be attached to this object. A blob can be removed from the object if it has a different displacement from the object. The difference between the object's bounding box (i.e. the rectangle circumscribing the object) and its prediction represents the correlation between the two. This difference can be in dimension or in position. If it exceeds 15% of the prediction region, the object's bounding box will be replaced by its prediction region. This percentage was estimated experimentally and works well for a large sample of video sequences. Fig. 6(a) and Fig. 6(b) depict the prediction region in a yellow box, the blobs in a green box, and the object in a dashed magenta box.

5 Experimental Results

This work aims at creating audio-video tools that will allow multimedia content producers to improve the richness of the multimedia experience for the blind, the deaf, the hard of hearing, and the hard of seeing, by automating key aspects of the multimedia production and post-production processes. In this context, the experiments are carried out on sequences extracted from two movies. Fig. 6, 7, and 8 illustrate the snapshots from 3 different video sequences. The sequences shown in Fig. 6 and 7 are extracted from the movie called "The Fabulous Destiny of Amélie Poulain". The sequence presented in Fig. 8 is extracted from the movie named "Life is a long quiet river". Of course, we cannot convey all the aspects of these sequences onto a static sheet of paper. Thus, only a few interesting frames were selected. The experiments reported in this section involve a mobile camera following an actor. The example reported in Fig. 6 depicts an actress walking from right to left, passing through the camera viewpoint. The camera follows the actress. Fig. 6(a) and 6(b) illustrate the shape of the object from several blobs. The green boxes indicate the blobs, the yellow box depicts the predicted area, the magenta box represents the object. Fig. 6(c) and 6(d) represent the combination of the forward tracking and the backward tracking process. The green box with label *Fdw* indicates the result of the forward tracking process and the blue box with label *Rev* indicates the result of the backward tracking process. Fig. 6(e) and 6(f) show the final result of tracking with the ID number on the top left of the box.

Fig. 7 shows a case where the camera follows an actor descending stairs and approaching the camera, then passing through and leaving the camera viewpoint. Note that the view of the actor is not lateral, but almost anterior. In Fig. 8, the camera stays focused on an actor who is moving away from the camera. Although these images are static they do express the dynamics of the whole sequence.

In order to validate the algorithm, ground-truth information on several video sequences is needed. Contrarily to standard video image databases, the ground-truth of sequences extracted from movies is not directly available and must be

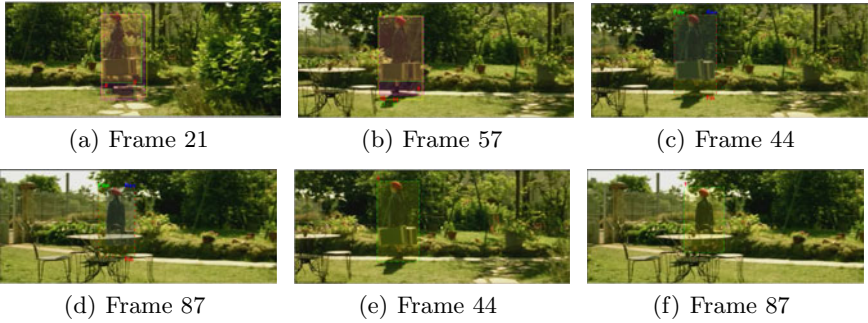


Fig. 6. Samples of tracking results (Sequence01)



Fig. 7. The snapshots of the tracking results of a video sequence extracted from the movie “The Fabulous Destiny of Amélie Poulain” (Sequence02)

generated manually. For each sequence, a program was used to display each frame on a monitor. Following, a human was asked to identify the object (i.e. the actor) and draw the bounding box around this object. All the bounding boxes are considered as the ground-truth of the sequence. In Fig. 7 and 8, the ground-truths are indicated by the red boxes, and the outputs of the algorithm are indicated by the green boxes. We observe that they are very close to each other.

The correspondence rate between these results and the ground-truth is then calculated to provide a better and more objective idea of the results obtained by the algorithm. The correspondence rate is defined as

$$C = 1 - \frac{N_N + N_P}{N_R + N_S} \quad (5)$$

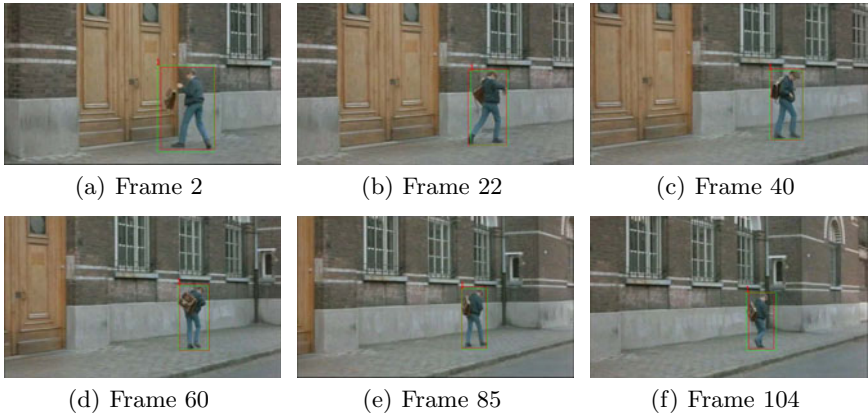


Fig. 8. Additional results of a video sequence extracted from the movie “Life is a long quiet river” (Sequence03)

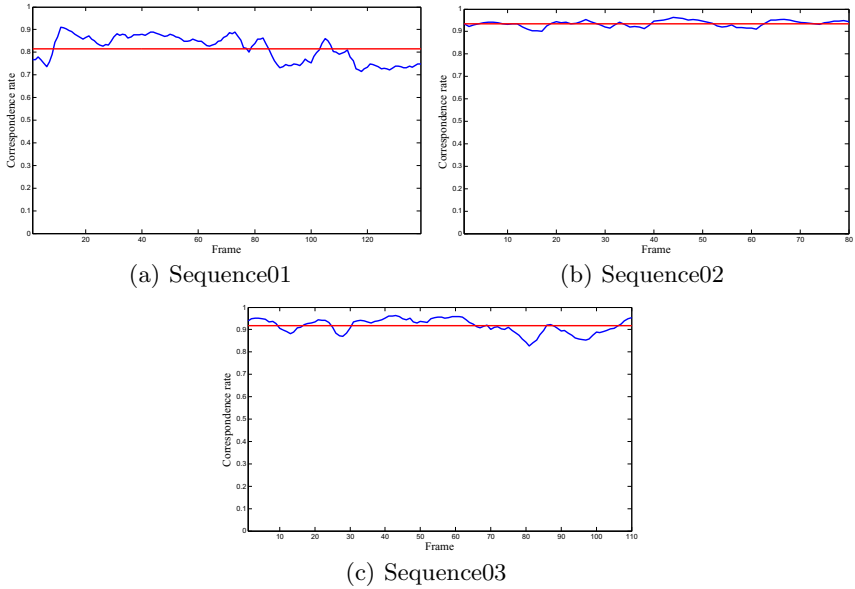


Fig. 9. Correspondence rate for the experimental sequences. Blue curve: correspondence rate between the results and the ground-truth. Red line: the average value of the correspondence rate.

where N_R is the total number of reference pixels in the ground-truth, N_S is the number of the object’s pixels. N_P is the number of false positives:

$$N_P = N_S - (N_S \cap N_R) \quad (6)$$

N_N is the number of false negatives:

$$N_N = N_R - (N_S \cap N_R) \quad (7)$$

The correspondence rate indicates how the output of the algorithm fits with the reference from the ground-truth. In other words, the higher the value of the correspondence rate between the results and the ground-truth, the better the performance of the algorithm.

The plots in Figure 9 show the correspondence rate between the outputs of our algorithm and the ground-truth. The average of the correspondence rate is represented as a red line, the correspondence rate is plotted in blue. According to these plots, the algorithm achieves the best results for Sequence02 with an average correspondence rate reaching 94% (Figure 9(b)). In Sequence01 (see Figure 9(a)), the average correspondence rate is about 81%. It is the lowest rate obtained among the experimental sequences. For other sequences, there are slight variations of the correspondence rate around its average value (see Figure 9(c) and 9(b)). These variations are caused by the cast shadows and false detections which are small and acceptable in the case of automatic video indexing. Moreover, the average correspondence rates obtained from these sequences are quite good (from 91% to 94%). Overall, we can conclude that the proposed algorithm works well on all of these extracted movie sequences.

6 Conclusion

We have proposed a method for tracking human motion in video sequences in which a moving camera is present. We can then reconstruct the 2D trajectories of the objects without requiring a high-level or special structural model. The results are promising, and the proposed method proves to be quite useful and convincing for video indexing. More experiments need to be conducted to validate the algorithm. Future work will involve conducting the experiments on additional scenarios with several actors, and with occlusions by other actors or scene components.

Acknowledgments. This work is financially supported by the E-Inclusion Network and the Department of Canadian Heritage through Canadian Culture Online.

References

1. Leung, M.K., Yang, Y.: A region based approach for human body motion analysis. *Pattern Recognition* 20(3), 321–339 (1987)
2. O'Rourke, J., Badler, N.I.: Model-Based image analysis of human motion using constraint propagation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 2(6), 522–536 (1980)
3. Ricquebourg, Y., Bouthemy, P.: Real-time tracking of moving persons by exploiting spatio-temporal image slices. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22(8), 797–808 (2000)

4. Polana, R., Nelson, R.: Low level recognition of human motion (or how to get your man without finding his body parts). In: Proceedings of the IEEE Workshop on Motion of Non-Rigid and Articulated Objects, pp. 77–82 (1994)
5. Xu, X., Li, B.: Rao-Blackwellised particle filter for tracking with application in visual surveillance. In: 2nd Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance 2005, pp. 17–24 (2005)
6. Zach, C., Pock, T., Bischof, H.: A duality based approach for realtime TV-L1 optical flow. In: Hamprecht, F.A., Schnörr, C., Jähne, B. (eds.) DAGM 2007. LNCS, vol. 4713, pp. 214–223. Springer, Heidelberg (2007)
7. Nguyen, N., Laurendeau, D., Branzan-Albu, A.: A robust method for camera motion estimation in movies based on optical flow. *International Journal of Intelligent Systems Technologies and Applications* 9, 228–238 (2010)
8. Otsu, N.: A threshold selection method from Gray-Level histograms. *IEEE Transactions on Systems, Man and Cybernetics* 9(1), 62–66 (1979)
9. Torresan, H., Turgeon, B., Ibarra-Castanedo, C., Hebert, P., Maldague, X.P., Burleigh, D.D., Cramer, K.E., Peacock, G.R.: Advanced surveillance systems: combining video and thermal imagery for pedestrian detection. In: *Thermosense XXVI*, Orlando, FL, USA, vol. 5405, pp. 506–515. SPIE, San Jose (April 2004)

Event Detection and Recognition Using Histogram of Oriented Gradients and Hidden Markov Models

Chun-hao Wang, Yongjin Wang, and Ling Guan

Ryerson University, Electrical and Computer Engineering, Toronto,
Ontario, Canada M5B 2K3
{cwang, ywang, lguan}@ee.ryerson.ca

Abstract. This paper presents an approach for object detection and event recognition in video surveillance scenarios. The proposed system utilizes a Histogram of Oriented Gradients (HOG) method for object detection, and a Hidden Markov Model (HMM) for capturing the temporal structure of the features. Decision making is based on the understanding of objects motion trajectory and the relationships between objects' movement and events. The proposed method is applied to recognize events from the public PETS and i-LIDS datasets, which include vehicle events such as U-turns and illegal parking, as well as abandoned luggage recognition established by set of rules. The effectiveness of the proposed solution is demonstrated through extensive experimentation.

Keywords: Activity recognition, object tracking, Hidden Markov Models, video surveillance, Histogram of Oriented Gradients.

1 Introduction

Recently, there has been extensive research on automatic video surveillance analysis, with potential application in biometrics, activity recognition, and human movement analysis. As a sub area of human computer interaction (HCI), the ability for machines to read and understand object and its movements is a much sought after goal. Advances in computer vision and image/video processing have made the processing of large amounts of data possible. Popular topics of research include biometrics, video surveillance, gesture recognition, and emotion recognition. These topics all look to create an intelligent system that can extract and process useful information from video sequences to improve automation and reduce man power required to process these overly abundant flow of video data.

Various approaches have been introduced in the literature on the topic of motion understanding and recognition. Du and Guan incorporated body shape features and Kalman filters to recognize humans' movement [1]. Hongeng et al. created a probabilistic finite automaton of event states from Bayesian networks for activity recognition [2].

Currently the PETS workshop has focused on video surveillance projects. Previous works have focused on various detection rules. Most of the proposed techniques for abandoned object detection rely on tracking information [7, 8, 9]. Lu et al. used Bayesian inference of context, spatial and temporal rules to determine abandoned packages [3]. It detects the time and position of luggage appearance, creating a combination of spatial and temporal rules. Smith et al. used Markov Chain Monte Carlo tracking and identify bags versus humans using likelihood functions of bag size and velocity [4]. Shet et al. used rule and language rules with Prolog for video monitoring [5]. The Prolog system can determine high level concepts such as thefts, illegal entry and unattended package with simple facts.

Current literature on video surveillance detection of special events of interest is usually based on heuristic rules. There are no formulations that can generalize to all security surveillance applications. Thus a probabilistic framework such as belief propagation using factor graphs can generalize and unite related concepts without specially created rules of thumb. Factor graphs and belief propagation has been used to infer high level semantic meaning by Naphade et al [6].

The task of detecting abandoned luggage is done by first segmenting objects of interest in the image sequence. If there are any objects detected, tracking is required to maintain identity and position of the objects. Given the objects are tracked correctly, special conditions may arise such as a luggage object left for an extended period of time, in which an alarm may be raised to signal the detection of a suspicious package for further handling.

The remainder of the paper is organized as follows. We discuss related works in section 2. The system and its technical details are described in section 3. We present the results in section 4 and end with concluding remarks in section 5.

2 Overview of the System

In this paper, we propose a novel solution to detect events in various surveillance video scenarios. Fig. 1 shows our system flowchart. The system includes four main components: 1) foreground and object detection using Gaussian mixture model based background subtraction and histogram of oriented gradients 2) computer object properties 3) input object movement trajectory to Hidden Markov Model 4) recognition of events based on movement of the object.

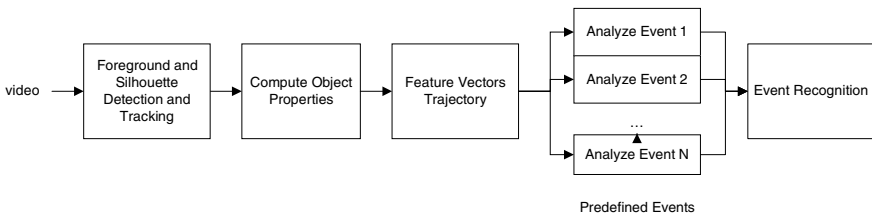


Fig. 1. System flowchart

2.1 Foreground Extraction

The first task of detecting objects in the image sequence requires a method for object extraction. Our system uses the Gaussian Mixture Model-based background subtraction method proposed in [7]. The distribution of each pixel is modeled as a Gaussian process, $\eta(\mu, \sigma)$,

The probability that an observed pixel will have an intensity value $I_{(x,y,t)}$ at (x, y) and time t is estimated by K Gaussian distributions as follows:

$$P(I_t) = \sum_{l=1}^K \frac{\omega_{l,t}}{(2\pi)^{1/2}} e^{-\frac{1}{2}(I_t - \mu_l)^T \Sigma_l^{-1} (I_t - \mu_l)}, \quad (1)$$

Where Σ_l is the Gaussian's covariance matrix, which is assumed to be diagonal. Thus the silhouette of objects is extracted if it is outside the range of the most likely pixel values:

$$I_{obj}(x, y, t) = \begin{cases} 0, & \text{if } |I_{(x,y,t)} - \mu_{(x,y,t)}| \leq \gamma \sigma_{(x,y,t)}, \\ 1, & \text{otherwise} \end{cases}, \quad (2)$$

where $I_{obj}(x, y, t)$ is the binary silhouette value of a pixel at (x, y) of the t^{th} frame, and γ is the determined threshold value. However, a static mean μ and standard deviation σ will not adapt to shifting backgrounds or changing lighting conditions. Thus given a location (x, y) , μ_t and ρ_t are updated using

$$\mu_t = (1 - \alpha)\mu_{t-1} + \alpha I_t \quad (3)$$

$$\sigma_t^2 = (1 - \alpha)\sigma_{t-1}^2 + \alpha(I_t - \mu_t)^T (I_t - \mu_t), \quad (4)$$

where α is the learning rate that determines the update speed of μ_t and σ_t . Adjusting γ changes the sensitivity of the silhouette detector, and $1/\alpha$ is proportional to time background is adjusted. Since PETS 2006 had detection rules of unattended packages being 30 seconds, $\alpha = 1/750$, $K = 3$, and $\gamma = 2$ is set empirically.

To enhance background subtraction, YUV colorspace is used instead of RGB. YUV and HSV are better than RGB as it more closely model the human perception of color changes. The pixels were not converted to grayscale but rather YUV Euclidean distance criteria can be used to find the silhouette,

$$\sqrt{(Y_t - \mu_{Y,t})^2 + (U_t - \mu_{U,t})^2 + (V_t - \mu_{V,t})^2} \leq \gamma \sigma_{(x,y,t)}, \quad (5)$$

which allows for better detection of color changes and less subtle illumination changes, which eliminates some shadows and reflections. Fig. 2 shows simple extraction results.

2.2 Camera-Real World Calibration

To compute useful physical characteristics of the objects in the surveillance video, it is assumed that cameras are stationary, and its location and view orientation are known. With extrinsic parameters of the camera (translation \mathbf{T} and rotation \mathbf{R})

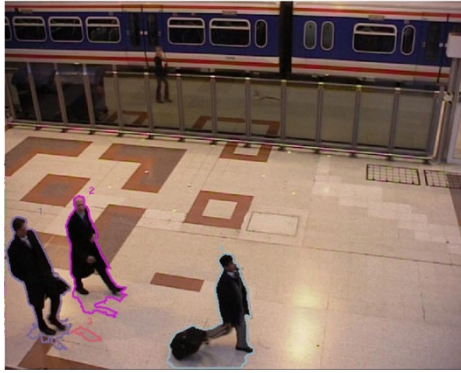


Fig. 2. Object extraction results



Fig. 3. Camera view transformation

provided in the PETS dataset, we calculate real world coordinates of objects from image coordinates using $\vec{P}_W = \vec{R}P_I + \vec{T}$, where P_W and P_I are world and image coordinate vectors (Fig. 3).

After converting to real world coordinates, spatio-temporal properties can be easily calculated and multi-view scenarios synchronized. All objects that are found to be standing outside of the area of interest can be discarded.

2.3 Histogram of Oriented Gradients

In our system, the HOG is used to detect foreground objects in a much higher accuracy than traditional methods. Dalal and Triggs [14] presented a general object detection algorithm with excellent detection results. Their method uses a dense grid of Histogram of Oriented Gradients, computed over blocks of size 16×16 pixels to represent a detection window. This representation proves to be powerful enough to classify humans using a linear Support Vector Machine (SVM). Dalal and Triggs used the single window that was successfully used for object representation [13].

HOG requires first a learning phase by grouping a normalized training data set and encoding images into features and learning the binary classifier. The sample HOG models are shown in Fig 4. After learning the object model, the steps for detecting starts by 1) scan the image at all scales and locations, 2) run classifier to obtain object/non-object decisions, and 3) fuse multiple detection in 3-D position and scale space.

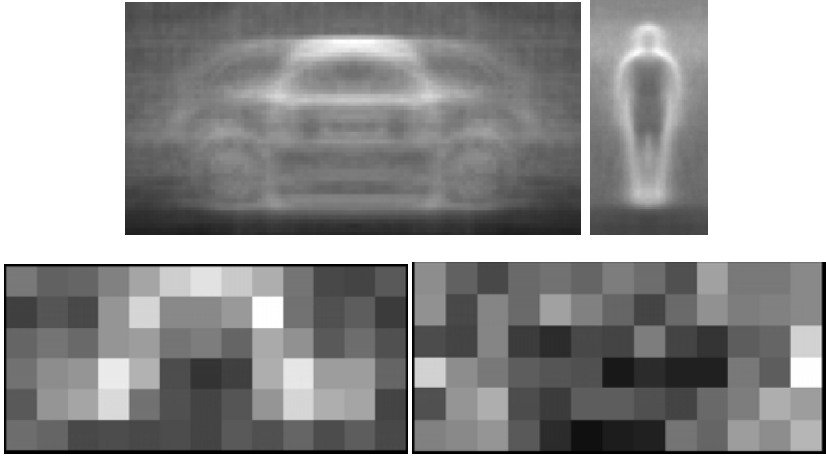


Fig. 4. A) HOG gradient model of a vehicle (sideview) and a human [13] B) Positive and negative weights from a trained car HOG model

The individual calculations of HOG is as follows: 1) Gamma compression 2) compute gradients 3) weighted vote in spatial & orientation cells 4) contrast normalize over overlapping spatial cells 5) collect HOGs over detection window 6) linear SVM. The creation process for HOG give it robust features in able to operate in multi-scale and multitude of object classes.

2.4 Hidden Markov Model (HMM)

To capture the statistical dependence across successive frames and identify the inherent temporal structure of the features, an HMM is employed to characterize the distribution of an image sequence of length N , which can be represented as $P(\omega_1, \omega_2, \dots, \omega_N)$. To this end, the resulting features of each frame are considered as the observation of an HMM [12], of which the probability density functions (PDF) given a state is modeled using Gaussian mixtures. To be specific, considering the case in which there are C classes, a feature sequence of length T of the k -th class, denoted as $W_k = [\omega_1^k, \omega_2^k, \dots, \omega_T^k]$, is used to train an N -state HMM, which can be represented using $\lambda_k = \{\pi_k, A_k, B_k\}$. Corresponding to the sequence of observations, we denote their respective hidden states as $Q_k = [q_1^k, q_2^k, \dots, q_T^k]$, each of which takes on the values of a finite set of states, denoted as $S_k = \{s_1, s_2, \dots, s_N\}$. Accordingly the first set of parameters of an HMM is composed of the initial state probabilities, i.e. $\pi_k = [\pi_1^k, \pi_2^k, \dots, \pi_N^k]$, where $\pi_n^k = P(q_0^k = s_i)$ and $i = 1, 2, \dots, N$. In addition, the second set of parameters consist of the state transition probabilities $A_k = [a_{ij}^k]_{N \times N}$, where $a_{ij}^k = P(q_t^k = s_j | q_{t-1}^k = s_i)$ and $i, j = 1, 2, \dots, N$. Finally, the third set of parameters $B_k = [p_1^k(\omega_t), p_2^k(\omega_t), \dots, p_N^k(\omega_t)]$ characterize the PDF's of a d -dimensional observation conditional on different states, which can be expressed as

$$p_i^k(\omega_t) = \sum_{m=0}^{M-1} P(\alpha_m^k) p(\omega_t | \alpha_m^k), \tag{6}$$

and

$$p(\omega_t | a_m^k) = \frac{e^{-\frac{1}{2}(\omega_t - \mu_m^k)^T (\Sigma_m^k)^{-1} (\omega_t - \mu_m^k)}}{(2\pi)^{d/2} |\Sigma_m^k|^{1/2}}, \quad (7)$$

Where μ_m^k and Σ_m^k are the mean vector and covariance matrix of the m -th mixture component of the k -th class. It should be noted that because we employ homogeneous HMMs the time indexes of the parameters can be dropped. The parameters can be estimated through the standard expectation maximization (EM) procedure.

Once the HMMs of all the classes are learned, the likelihood values of a new audio or video sample $W = [\omega_1, \omega_2, \dots, \omega_T]$ with respect to different classes can be calculated through

$$\begin{aligned} p &= (\omega_1, \omega_2, \dots, \omega_T) \quad (8) \\ &= \sum_{q_1^k} \dots \sum_{q_T^k} P(\omega_1^k, \omega_2^k, \dots, \omega_T^k; q_1^k, q_2^k, \dots, q_T^k,) \\ &= \sum_{q_1^k} \dots \sum_{q_T^k} P(q_0^k) \prod_{t=1}^T p_t(\omega_t^k) \prod_{t=2}^T P(q_t^k | q_{t-1}^k). \end{aligned}$$

The extracted foreground objects' center of gravity can produce a traced trajectory is input into the HMM. The discontinuity of tracks of moving objects often arises when moving objects are not detected for a few frames, such as from total occlusion, or when all regions do not satisfy the ground plan assumption. In this case, hypotheses about connections of fragments of tracks are smoothed over a few frames. The output of the HMM contains the likelihood of the query sample with respect to difference classes. In our system, the number of hidden states was chosen to be three. The HMM classifier requires no parameter tuning so it is relatively cheap in terms of concept detection performance.

3 Experimental Results

We tested our approach with the PETS2006 [11] and i-LIDS [10] datasets, which was designed to contain real world video surveillance footage in a public environment. The ground truth for the sequences includes the type of objects and also the coordinates and time of events of interest. The HOG is trained for a vehicle top view and an upright standing human trained from the PETS2006 and i-LIDS datasets.

PETS2006 dataset contains 4 camera multi-sensor sequences containing left-luggage scenarios with increasing complexity. There are seven scenarios each with 4 different capture view points. Our algorithm processes one camera viewpoint at a time, and the results are from the camera view where objects are clearer and bigger. Table 1 and Fig. 3 show a summary of the dataset broken down into details.

Table 1. Challenges in the PETS 2006 dataset

Seq.	Length (s)	Luggage items	People nearby	Abandoned	Difficulty (PETS)
S1	121	1 backpack	1	Yes	1/5
S2	102	1 suitcase	2	Yes	3/5
S3	94	1 briefcase	1	No	1/5
S4	122	1 suitcase	2	Yes	4/5
S5	136	1 ski	1	Yes	2/5
S6	112	1 backpack	2	Yes	3/5
S7	136	1 suitcase	6	Yes	5/5

Table 2. PETS2006 Luggage and Alarm Detection Results

Seq.			Luggage Detection	Loc (x, y)	Alarm	Alarm Time
S1	ground	truth	Yes	(.22, -.44)	Yes	113.7s
	result		Yes	(.22, -.34)	Yes	113.2s
	error		0%	0.10m	0%	0.5s
S2	ground	truth	Yes	(.34, -.52)	Yes	91.8s
	result		Yes	(.22, -.33)	Yes	90.8s
	error		0%	0.20m	0%	1.08s
S3	ground	truth	Yes	(.86, -.54)	No	-
	result		No	-	No	-
	error		100%	-	0%	-
S4	ground	truth	Yes	(.24, -.27)	Yes	104.1s
	result		No	(.10, -.03)	No	-
	error		0%	0.25m	100%	-
S5	ground	truth	Yes	(.34, -.56)	Yes	110.6s
	result		Yes	(.24, -.49)	Yes	110.6s
	error		0%	0.13m	0%	0.0s
S6	ground	truth	Yes	(.80, -.78)	Yes	96.9s
	result		Yes	(.69, -.49)	Yes	96.9s
	error		0%	0.30m	0%	0.0s
S7	ground	truth	Yes	(.35, -.57)	Yes	94.0s
	result		Yes	(.30, -.34)	Yes	90.4s
	error		0%	0.23m	0%	3.6s

Tables 2 and 3 show our results for PETS2006, with 6 out of 7 sequences successfully detecting an abandoned luggage and raising an alarm. The alarm time and location of the events can be compared to fine-tune the results, and our results show that there are usually a few frames of delay between the ground truth data, which can attribute up to 0.30 meters of error in real-world coordinates.

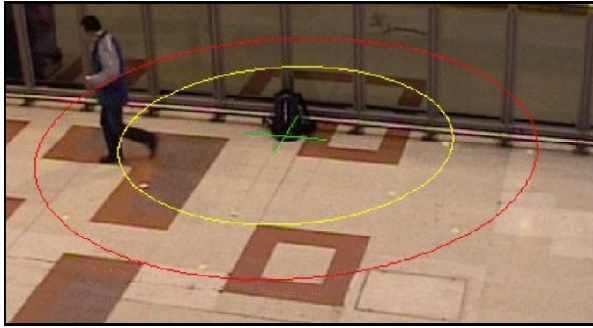


Fig. 5. Alarm conditions. Owner outside of red ring (3 meters) represents abandoned luggage, between yellow (2 meters) and red represents unattended luggage, and green cross represents the position of the bag.

i-LIDS dataset (Fig. 6) contains Quicktime MJPEG real CCTV footage based on four scenarios: 1) Parked vehicles 2) Abandoned baggage 3) Sterile Zone and 4) Doorway surveillance. Within the scenarios, certain alarm events are defined, for example, the presence of a parked vehicle in a defined zone for more than 60 seconds. This alarm can represent an illegally parked vehicle, and is an event of interest.



Fig. 6. i-LIDS dataset, vehicle surveillance scenarios

Table 4 and 5 shows the results of our method performed on i-LIDS dataset. The two main events are abandoned luggage in a train station and vehicle parking surveillance in a street. The results show that all abandoned luggage were detected, but with 3 false positives, generated from static people misidentified as an abandoned object.

Table 3. Abandoned object detection for PETS2006

# of sequences	abandoned objects	True Positives	False Positives
7	7	6	0

Table 4. Abandoned object detection for i-Lids dataset

# of sequences	Abandoned objects	True Positives	False Positives
5	8	8	3

Table 5. Illegally Parked vehicle detection for i-Lids dataset parked vehicle scenario

# of sequences	Parked vehicles	True Positives	False Positives
5	6	6	1

4 Conclusion and Future Work

The detection of abandoned packages is an important application in video surveillance and security. In this paper a framework for event detection using HMM and HoG was demonstrated to detect abandoned packages. Results show that our method is accurate in detecting, tracking, and recognizing 3 different scenarios across datasets. In certain circumstances, there are false positives generated from misclassifying a static person as an abandoned luggage. For purposes of video surveillance and detection of suspicious events, false positives' costs can be much lower than a false negative.

Future research direction can apply our framework towards other surveillance and activity and event recognition tasks. Possibilities include determining other suspicious behaviors, thefts, loitering, as well as home based use such as helper surveillance in elderly homes and hospitals, and finally more complex scenarios involving multiple views and scenarios involving multiple actors.

The experiments show promising results with different possible future research directions, such as only being monocular, and HOGs require object models trained according to the scenario type. Future research can further improve the robustness of our method by having automatic adaptation to normal scenes and suspicious events automatically. Our foreground extraction method can also be more robust, to account for excessive motion and constant lighting changes. The extensive testing results proved that our approach can be applied to real-world surveillance scenarios.

References

1. Du, M., Guan, L.: Human recognition by body shape features. Proc. of SPIE-IS&T Electronic Imaging, pp. 535–544 (2005)
2. Hongeng, S., Nevatia, R., Bremond, F.: Video-based event recognition: activity representation and probabilistic recognition methods. In: Computer Vision and Image Understanding CVPR, pp. 129–162 (2004)
3. Lu S., Zhang J., Fend D.D.: Detecting unattended packages through human activity recognition and object association. Pattern Recognition Society, 2173–2184 (2007)

4. Smith, K., Quelhas, P., Gatica-Perez, D.: Detecting abandoned luggage items in a public space. In: Proc. IEEE International Workshop on PETS, pp. 75–82 (2006)
5. Shet, V.D., Harwood, D., Davis, L.S.: VidMAP: Video Monitoring of Activity with Prolog. In: IEEE International Conference on Advanced Video and Signal-based Surveillance, pp. 224–229 (2005)
6. Naphade, M.R., Kozintsev, I.V., Huang, T.H.: A Factor Graph Framework for Semantic Video Indexing. *IEEE Transactions on Circuits and Systems for Video Technology*, 40–52 (2002)
7. Stauffer, C., Grimson, W.E.L.: Adaptive background mixture models for real-time tracking. In: IEEE Conference on Computer Vision and Pattern Recognition, pp. 40–52 (1998)
8. Belongie, J.M.S., Puzicha, J.: Shape matching object recognition using shape contexts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 509–522 (2002)
9. Breitenstein, M.D., Reichlin, F., Leibe, B., Koller-Meier, E., Gool, L.V.: Robust tracking-by-detection using a detector confidence particle filter. In: ICCV (2009)
10. i-LIDS dataset for AVSS (2007),
<ftp://motinas.elec.qmul.ac.uk/pub/iLids/>
11. PETS 2006, Benchmark Data (2006),
<http://www.cvg.rdg.ac.uk/PETS2006/data.html>
12. Rabiner, L.R.: A tutorial on Hidden Markov Models and selected applications in speech recognition. *Proceedings of the IEEE* 77(2), 257–286 (1989)
13. Dalal, N., Triggs, B.: Object Detection using Histograms of Oriented Gradients. In: Pascal VOC Workshop, ECCV (2006)
14. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: Conference on Computer Vision and Pattern Recognition (2005)

Author Index

- Abdoola, Rishaad II-317
Ahmad, Irfan II-397
Ahmadi, Majid I-69
Akgul, Yusuf Sinan I-304
Al-Khatib, Wasfi G. II-397
Allili, Mohand Saïd I-314
Almeida, João Dallyson S. II-151
Alonso, Luis II-360
Alshayeb, Mohammed II-397
Ammar, Moez II-348
An, Huiyao II-89
Armato, Samuel G. II-21
Asari, Vijayan K. I-30
Asraf, Daniel II-101
Ayatollahi, Ahmad II-48
Aziz, Kheir-Eddine II-170
- Baja, Gabriella Sanniti di I-344
Bakina, Irina II-130
Bao, Huiyun I-262
Bedawi, Safaa M. II-307
Bernardino, Alexandre I-294
Bhatnagar, Gaurav II-286
Bouguila, Nizar I-201
Branzan-Albu, Alexandra I-426
Brun, Luc I-173
Brunet, Dominique I-100, II-264
Burke, Robert D. II-12
Busch, Andrew II-389
- Campilho, Aurélio II-1, II-68
Cancela, B. I-416
Candemir, Sema I-304
Carmona, Pedro Latorre II-360
Chae, Oksam I-274
Chen, Cunjian II-120
Cheng, Howard II-243
Clark, Adrian F. I-253
Conte, Donatello I-173
Cordes, Kai I-161
Cordier, Frédéric I-365
Cunha, João Paulo Silva II-59
- Dahmane, Mohamed II-233
Dai, Xiaochen I-395
- Das, Sukhendu II-212
Dastmalchi, Hamidreza I-193
Debayle, Johan I-183
Dechev, Nikolai II-12
Dewitte, Walter II-1
Ding, Yan II-276
Djouani, Karim I-80
Driessen, P.F. II-328
Du, Shengzhi I-80
Duric, Zoran I-221
- Ehsan, Shoaib I-253
Elguebaly, Tarek I-201
Esmailsabzali, Hadi II-12
- Faez, Karim I-193, II-161
Fernandes, José Maria II-59
Fernández, A. I-416
Ferrari, Giselle II-40
Fertil, Bernard II-170
Fieguth, Paul I-385
Figueira, Dario I-294
Fleck, Daniel I-221
Foggia, Pasquale I-173
Frejlichowski, Dariusz II-380
Furst, Jacob II-21
- Gangeh, Mehrdad J. I-335
Gao, Meng I-406
George, Loay E. II-253
Ghodsí, Ali I-335
Guan, Ling I-436, II-79, II-111, II-140
Gupta, Rachana A. II-338
- Hamam, Yskandar I-80
Hancock, Edwin II-89
Hardeberg, Jon Y. I-375
Hassen, Rania I-40
Hégarat-Masclé, Sylvie Le II-348
Homola, Ondřej II-31
Huang, Jiawei I-122
Huang, Lei II-222
Huang, Xiaozheng II-276
- Ibrahim, Muhammad Talal II-79, II-111

- Kadim, Azhar M. II-253
 Kamel, Mohamed S. I-335, I-385, II-307
 Kang, Yousun I-141
 Kanwal, Nadia I-253
 Khan, M. Aurangzeb II-79
 Klempt, Carsten I-161
 Konvalinka, Ira II-101
 Krylov, Andrey S. I-284
 Kumazawa, Itsuo I-21
 Kurakin, Alexey II-130
- Laurendeau, Denis I-426
 Le, Tam T. I-141
 Leboeuf, Karl I-69
 Li, Qi I-232
 Li, Xueqing I-262
 Li, Ze-Nian I-122
 Liang, Jie II-276
 Lisowska, Agnieszka I-50
 Liu, Bin I-90
 Liu, Changping II-222
 Liu, Huaping I-406
 Liu, Jiangchuan II-276
 Liu, Wei I-325
 Liu, Weijie I-90
 Liu, Ying II-370
 Luong, Hiệp Q. I-11
- Mahmoud, Sabri A. II-397
 Maillot, Yvan I-183
 Makaremi, Iman I-69
 Mandava, Ajay K. I-58
 Mansouri, Alamin I-375
 Marsico, Maria De II-191
 McClean, Sally I-211
 McDonald-Maier, Klaus D. I-253
 Mehmood, Tariq II-79
 Melkemi, Mahmoud I-365
 Mendonça, Ana Maria II-1, II-68
 Merad, Djamel II-170
 Mestetskiy, Leonid II-130
 Meunier, Jean II-233
 Miao, Yun-Qian I-385
 Mizotin, Maxim M. I-284
 Moan, Steven Le I-375
 Momani, Bilal Al I-211
 Monacelli, Eric II-317
 Moreno, Jose E. II-360
 Moreno, Plinio I-152
 Morrow, Philip I-211
- Mounier, Hugues II-348
 Murray, Jim II-1
 Murshed, Mahbub I-274
- Nappi, Michele II-191
 Nguyen, Nhat-Tan I-426
 Nguyen, Thuc D. I-141
 Nicolo, Francesco II-180
 Nieuwland, Jeroen II-1
 Noel, Guillaume II-317
- Ortega, M. I-416
 Oskuie, Farhad Bagher II-161
 Ostermann, Jörn I-161
- Paiva, Anselmo C. II-151
 Palk, Phillip II-389
 Park, Edward J. II-12
 Paulhac, Ludovic I-354
 Payandeh, Shahram I-395
 Pedrocca, Pablo Julian I-314
 Penedo, Manuel G. I-416
 Percannella, G. II-297
 Petrou, Maria I-132
 Philips, Wilfried I-11
 Pinoli, Jean-Charles I-183
 Pizurica, Aleksandra I-11
 Pla, Filiberto II-360
 Presles, Benoît I-183
- Quddus, Azhar II-101
 Quelhas, Pedro II-1
 Quivy, Charles-Henri I-21
- Raicu, Daniela S. II-21
 Raman, Balasubramanian II-286
 Ramel, Jean-Yves I-354
 Ramirez, Adin I-274
 Regentova, Emma E. I-58
 Renard, Tom I-354
 Ribeiro, Eraldo I-325
 Ribeiro, Pedro I-152
 Riccio, Daniel II-191
 Rosenhahn, Bodo I-161
 Ross, Arun II-120
 Rudrani, Shiva II-212
 Ružić, Tijana I-11
- Sakaki, Kelly II-12
 Salama, Magdy I-40

- Sá-Miranda, M. Clara II-68
 Santhaseelan, Varun I-30
 Santos-Victor, José I-152
 Sapidis, Nickolas S. I-365
 Sattar, F. II-328
 Schaaf, Crystal II-360
 Scherer, Manuel I-161
 Schmid, Natalia A. II-180
 Serino, Luca I-344
 Shafie, Siti Mariam I-132
 Siena, Stephen A. II-21
 Silva, Aristófanos C. II-151
 Snyder, Wesley E. II-338
 Song, Caifang II-201
 Sorokin, Dmitry V. I-284
 Sousa, António V. II-68
 Stejskal, Stanislav II-31
 Sugimoto, Akihiro I-141
 Sun, Fuchun I-406
 Sun, Yanfeng II-201
 Svoboda, David II-31

 Tafula, Sérgio II-59
 Talebi, Mohammad II-48
 Toda, Sorin II-101
 Topic, Oliver I-161
 Tran, Son T. I-141
 Trivedi, Vivek II-243
 Tu, Chunling I-80
 Tzanetakis, G. II-328

 Venetsanopoulos, A.N. II-111, II-140
 Vento, M. II-297

 Vento, Mario I-173
 Voisin, Yvon I-375
 Vrscaj, Edward R. I-100, II-264

 Wang, Chun-hao I-436
 Wang, Lei I-242
 Wang, Yongjin I-436, II-111, II-140
 Wang, Zemin II-370
 Wang, Zhaozhong I-242
 Wang, Zhou I-1, I-40, I-100,
 I-111, II-264
 Wechsler, Harry II-191
 Wu, Q.M. Jonathan II-286
 Wyk, Barend Jacobus van I-80, II-317

 Xiong, Pengfei II-222
 Xu, Tao II-370

 Yang, Wen II-370
 Yano, Vitor II-40
 Yeganeh, Hojatollah I-111
 Yin, Baocai II-201
 Yin, Jianping II-89

 Zeng, Kai I-1
 Zhang, Jianming II-89
 Zhang, Rui II-140
 Zhang, Ziming I-122
 Zhou, Chunxia II-370
 Zhu, En II-89
 Zimmer, Alessandro II-40
 Zinovev, Dmitriy II-21
 Zinoveva, Olga II-21