

Kenji Suzuki
Fei Wang
Dinggang Shen
Pingkun Yan (Eds.)

LNCS 7009

Machine Learning in Medical Imaging

Second International Workshop, MLMI 2011
Held in Conjunction with MICCAI 2011
Toronto, Canada, September 2011, Proceedings



Springer

Commenced Publication in 1973

Founding and Former Series Editors:

Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

David Hutchison

Lancaster University, UK

Takeo Kanade

Carnegie Mellon University, Pittsburgh, PA, USA

Josef Kittler

University of Surrey, Guildford, UK

Jon M. Kleinberg

Cornell University, Ithaca, NY, USA

Alfred Kobsa

University of California, Irvine, CA, USA

Friedemann Mattern

ETH Zurich, Switzerland

John C. Mitchell

Stanford University, CA, USA

Moni Naor

Weizmann Institute of Science, Rehovot, Israel

Oscar Nierstrasz

University of Bern, Switzerland

C. Pandu Rangan

Indian Institute of Technology, Madras, India

Bernhard Steffen

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

Kenji Suzuki Fei Wang Dinggang Shen
Pingkun Yan (Eds.)

Machine Learning in Medical Imaging

Second International Workshop, MLMI 2011
Held in Conjunction with MICCAI 2011
Toronto, Canada, September 18, 2011
Proceedings

Volume Editors

Kenji Suzuki
The University of Chicago
Chicago, IL 60637, USA
E-mail: suzuki@uchicago.edu

Fei Wang
IBM Research Almaden
San Jose, CA 95120, USA
E-mail: wangfe@us.ibm.com

Dinggang Shen
University of North Carolina
Chapel Hill, NC 27510, USA
E-mail: dgshen@med.unc.edu

Pingkun Yan
Chinese Academy of Sciences
Xian Institute of Optics and Precision Mechanics
Xi'an, Shaanxi 710119, China
E-mail: pingkun.yan@opt.ac.cn

ISSN 0302-9743

ISBN 978-3-642-24318-9

DOI 10.1007/978-3-642-24319-6

Springer Heidelberg Dordrecht London New York

e-ISSN 1611-3349

e-ISBN 978-3-642-24319-6

Library of Congress Control Number: 2011936535

CR Subject Classification (1998): I.4, I.5, J.3, I.2, I.2.10, I.3.3

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

The Second International Workshop on Machine Learning in Medical Imaging (MLMI) 2011 was held at Westin Harbour Castle, Toronto, Canada, on September 18, 2011 in conjunction with the 14th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI).

Machine learning plays an essential role in the medical imaging field, including computer-aided diagnosis, image segmentation, image registration, image fusion, image-guided therapy, image annotation and image database retrieval. With advances in medical imaging, new imaging modalities and methodologies—such as cone-beam/multi-slice CT, 3D ultrasound imaging, tomosynthesis, diffusion-weighted MRI, positron-emission tomography (PET)/CT, electrical impedance tomography and diffuse optical tomography—as well as new machine-learning algorithms/applications are demanded in the medical imaging field. Single-sample evidence provided by the patient’s imaging data is often not sufficient to provide satisfactory performance. Because of large variations and complexity, it is generally difficult to derive analytic solutions or simple equations to represent objects such as lesions and anatomy in medical images. Therefore, tasks in medical imaging require learning from examples for accurate representation of data and prior knowledge.

MLMI 2011 was the second in a series of workshops on this topic. The main aim of this workshop is to help advance scientific research within the broad field of machine learning in medical imaging. This workshop focuses on major trends and challenges in this area, and it presents work aimed at identifying new cutting-edge techniques and their use in medical imaging. We hope the series of workshops becomes a new platform for translating research from the bench to the bedside.

The range and level of submissions for this year’s meeting were of very high quality. Authors were asked to submit full-length papers for review. A total of 74 papers were submitted to the workshop in response to the call for papers. Each of the 74 papers underwent a rigorous double-blinded peer-review process, with each paper being reviewed by at least two (typically three) reviewers in the Program Committee composed of 50 known experts in the field. Based on the reviewing scores and critiques, the 44 best papers (59%) were accepted for presentation at the workshop and chosen to be included in this Springer LNCS volume. The large variety of machine-learning techniques necessary for and applied to medical imaging was well represented at the workshop.

We are grateful to the Program Committee for reviewing submitted papers and giving constructive comments and critiques, to authors for submitting high-quality papers, to presenters for excellent presentations, and to all those who supported MLMI 2011 by attending the meeting.

July 2011

Kenji Suzuki
Fei Wang
Dinggang Shen
Pingkun Yan

Organization

Program Committee

David Beymer	IBM Research, USA
Guangzhi Cao	GE Healthcare, USA
Heang-Ping Chan	University of Michigan Medical Center, USA
Sheng Chen	University of Chicago, USA
Zohara Cohen	NIBIB, NIH, USA
Marleen de Bruijne	University of Copenhagen, Denmark
Yong Fan	Chinese Academy of Sciences, China
Roman Filipovych	University of Pennsylvania, USA
Alejandro Frangi	Pompeu Fabra University, Spain
Hayit Greenspan	Tel Aviv University, Israel
Ghassan Hamarneh	Simon Fraser University, Canada
Joachim Hornegger	Friedrich Alexander University, Germany
Steve Jiang	University of California, San Diego, USA
Xiaoyi Jiang	University of Münster, Germany
Nico Karssemeijer	Radboud University Nijmegen Medical Centre, The Netherlands
Minjeong Kim	University of North Carolina, Chapel Hill, USA
Ritwik Kumar	IBM Almaden Research Center, USA
Shuo Li	GE Healthcare, Canada
Yang Li	Allen Institute for Brain Science, USA
Marius Linguraru	National Institutes of Health, USA
Yoshitaka Masutani	University of Tokyo, Japan
Marc Niethammer	University of North Carolina, Chapel Hill, USA
Ipek Oguz	University of North Carolina, Chapel Hill, USA
Kazunori Okada	San Francisco State University, USA
Sebastien Ourselin	University College London, UK
Kilian M. Pohl	University of Pennsylvania, USA
Yu Qiao	Shanghai Jiao Tong University, China
Xu Qiao	University of Chicago, USA
Daniel Rueckert	Imperial College London, UK
Clarisa Sanchez	University Medical Center Utrecht, The Netherlands
Li Shen	Indiana University School of Medicine, USA
Akinobu Shimizu	Tokyo University of Agriculture and Technology, Japan
Min C. Shin	University of North Carolina, Charlotte, USA
Hotaka Takizawa	University of Tsukuba, Japan
Xiaodong Tao	GE Global Research, USA

VIII Organization

Bram van Ginneken	Radboud University Nijmegen Medical Centre, The Netherlands
Axel W.E. Wismueller	University of Rochester, USA
Guorong Wu	University of North Carolina, Chapel Hill, USA
Jianwu Xu	University of Chicago, USA
Yiqiang Zhan	Siemens Medical Solutions, USA
Daoqiang Zhang	Nanjing University of Aeronautics and Astronautics, China
Yong Zhang	IBM Almaden Research Center, USA
Bin Zheng	University of Pittsburgh, USA
Guoyan Zheng	University of Bern, Switzerland
Kevin Zhou	Siemens Corporate Research, USA
Sean Zhou	Siemens Medical Solutions, USA
Xiangrong Zhou	Gifu University, Japan
Luping Zhou	CSIRO, Australia
Yun Zhu	University of California, San Diego, USA
Hongtu Zhu	University of North Carolina, Chapel Hill, USA

Table of Contents

Learning Statistical Correlation of Prostate Deformations for Fast Registration	1
<i>Yonghong Shi, Shu Liao, and Dinggang Shen</i>	
Automatic Segmentation of Vertebrae from Radiographs: A Sample-Driven Active Shape Model Approach	10
<i>Peter Mysling, Kersten Petersen, Mads Nielsen, and Martin Lillholm</i>	
Computer-Assisted Intramedullary Nailing Using Real-Time Bone Detection in 2D Ultrasound Images	18
<i>Agnès Masson-Sibut, Amir Nakib, Eric Petit, and François Leitner</i>	
Multi-Kernel Classification for Integration of Clinical and Imaging Data: Application to Prediction of Cognitive Decline in Older Adults ...	26
<i>Roman Filipovych, Susan M. Resnick, and Christos Davatzikos</i>	
Automated Selection of Standardized Planes from Ultrasound Volume	35
<i>Bahbibi Rahmatullah, Aris Papageorghiou, and J. Alison Noble</i>	
Maximum Likelihood and James-Stein Edge Estimators for Left Ventricle Tracking in 3D Echocardiography	43
<i>Engin Dikici and Fredrik Orderud</i>	
A Locally Deformable Statistical Shape Model	51
<i>Carsten Last, Simon Winkelbach, Friedrich M. Wahl, Klaus W.G. Eichhorn, and Friedrich Bootz</i>	
Monte Carlo Expectation Maximization with Hidden Markov Models to Detect Functional Networks in Resting-State fMRI	59
<i>Wei Liu, Suyash P. Awate, Jeffrey S. Anderson, Deborah Yurgelun-Todd, and P. Thomas Fletcher</i>	
DCE-MRI Analysis Using Sparse Adaptive Representations	67
<i>Gabriele Chiusano, Alessandra Staglioniò, Curzio Basso, and Alessandro Verri</i>	
Learning Optical Flow Propagation Strategies Using Random Forests for Fast Segmentation in Dynamic 2D & 3D Echocardiography	75
<i>Michael Verhoeck, Mohammad Yaqub, John McManigle, and J. Alison Noble</i>	

A Non-rigid Registration Framework That Accommodates Pathology Detection	83
<i>Chao Lu and James S. Duncan</i>	
Segmentation Based Features for Lymph Node Detection from 3-D Chest CT	91
<i>Johannes Feulner, S. Kevin Zhou, Matthias Hammon, Joachim Hornegger, and Dorin Comaniciu</i>	
Segmenting Hippocampus from 7.0 Tesla MR Images by Combining Multiple Atlases and Auto-Context Models	100
<i>Minjeong Kim, Guorong Wu, Wei Li, Li Wang, Young-Don Son, Zang-Hee Cho, and Dinggang Shen</i>	
Texture Analysis by a PLS Based Method for Combined Feature Extraction and Selection	109
<i>Joselene Marques and Erik Dam</i>	
An Effective Supervised Framework for Retinal Blood Vessel Segmentation Using Local Standardisation and Bagging	117
<i>Uyen T.V. Nguyen, Alauddin Bhuiyan, Kotagiri Ramamohanarao, and Laurence A.F. Park</i>	
Automated Identification of Thoracolumbar Vertebrae Using Orthogonal Matching Pursuit	126
<i>Tao Wu, Bing Jian, and Xiang Sean Zhou</i>	
Segmentation of Skull Base Tumors from MRI Using a Hybrid Support Vector Machine-Based Method	134
<i>Jiayin Zhou, Qi Tian, Vincent Chong, Wei Xiong, Weimin Huang, and Zhimin Wang</i>	
Spatial Nonparametric Mixed-Effects Model with Spatial-Varying Coefficients for Analysis of Populations	142
<i>Juan David Ospina, Oscar Acosta, Gaël Drán, Guillaume Cazoulat, Antoine Simon, Juan Carlos Correa, Pascal Haigron, and Renaud de Crevoisier</i>	
A Machine Learning Approach to Tongue Motion Analysis in 2D Ultrasound Image Sequences	151
<i>Lisa Tang, Ghassan Hamarneh, and Tim Bressmann</i>	
Random Forest-Based Manifold Learning for Classification of Imaging Data in Dementia	159
<i>Katherine R. Gray, Paul Aljabar, Rolf A. Heckemann, Alexander Hammers, and Daniel Rueckert</i>	
Probabilistic Graphical Model of SPECT/MRI	167
<i>Stefano Pedemonte, Alexandre Bousse, Brian F. Hutton, Simon Arridge, and Sébastien Ourselin</i>	

Directed Graph Based Image Registration	175
<i>Hongjun Jia, Guorong Wu, Qian Wang, Yaping Wang, Minjeong Kim, and Dinggang Shen</i>	
Improving the Classification Accuracy of the Classic RF Method by Intelligent Feature Selection and Weighted Voting of Trees with Application to Medical Image Segmentation	184
<i>Mohammad Yaqub, M. Kassim Javaid, Cyrus Cooper, and J. Alison Noble</i>	
Network-Based Classification Using Cortical Thickness of AD Patients	193
<i>Dai Dai, Huiguang He, Joshua Vogelstein, and Zengguang Hou</i>	
Anatomical Regularization on Statistical Manifolds for the Classification of Patients with Alzheimer's Disease	201
<i>Rémi Cuingnet, Joan Alexis Glaunès, Marie Chupin, Habib Benali, and Olivier Colliot</i>	
Rapidly Adaptive Cell Detection Using Transfer Learning with a Global Parameter	209
<i>Nhat H. Nguyen, Eric Norris, Mark G. Clemens, and Min C. Shin</i>	
Automatic Morphological Classification of Lung Cancer Subtypes with Boosting Algorithms for Optimizing Therapy	217
<i>Ching-Wei Wang and Cheng-Ping Yu</i>	
Hot Spots Conjecture and Its Application to Modeling Tubular Structures	225
<i>Moo K. Chung, Seongho Seo, Nagesh Adluru, and Houri K. Vorperian</i>	
Fuzzy Statistical Unsupervised Learning Based Total Lesion Metabolic Activity Estimation in Positron Emission Tomography Images	233
<i>Jose George, Kathleen Vunckx, Sabine Teijpar, Christophe M. Deroose, Johan Nuyts, Dirk Loeckx, and Paul Suetens</i>	
Predicting Clinical Scores Using Semi-supervised Multimodal Relevance Vector Regression	241
<i>Bo Cheng, Daoqiang Zhang, Songcan Chen, and Dinggang Shen</i>	
Automated Cephalometric Landmark Localization Using Sparse Shape and Appearance Models	249
<i>Johannes Keustermans, Dirk Smeets, Dirk Vandermeulen, and Paul Suetens</i>	
A Comparison Study of Inferences on Graphical Model for Registering Surface Model to 3D Image	257
<i>Yoshihide Sawada and Hidekata Hontani</i>	

A Large-Scale Manifold Learning Approach for Brain Tumor Progression Prediction	265
<i>Loc Tran, Deb Banerjee, Xiaoyan Sun, Jihong Wang, Ashok J. Kumar, David Vinning, Frederic D. McKenzie, Yaohang Li, and Jiang Li</i>	
Automated Detection of Major Thoracic Structures with a Novel Online Learning Method	273
<i>Nima Tajbakhsh, Hong Wu, Wenzhe Xue, and Jianming Liang</i>	
Accurate Regression-Based 4D Mitral Valve Surface Reconstruction from 2D+t MRI Slices	282
<i>Dime Vitanovski, Alexey Tsymbal, Razvan Ioan Ionasec, Michaela Schmidt, Andreas Greiser, Edgar Mueller, Xiaoquang Lu, Gareth Funka-Lea, Joachim Hornegger, and Dorin Comaniciu</i>	
Tree Structured Model of Skin Lesion Growth Pattern via Color Based Cluster Analysis	291
<i>Sina Khakabi, Tim K. Lee, and M. Stella Atkins</i>	
Subject-Specific Cardiac Segmentation Based on Reinforcement Learning with Shape Instantiation	300
<i>Lichao Wang, Su-Lin Lee, Robert Merrifield, and Guang-Zhong Yang</i>	
Faster Segmentation Algorithm for Optical Coherence Tomography Images with Guaranteed Smoothness	308
<i>Lei Xu, Branislav Stojkovic, Hu Ding, Qi Song, Xiaodong Wu, Milan Sonka, and Jinhui Xu</i>	
Automated Nuclear Segmentation of Coherent Anti-Stokes Raman Scattering Microscopy Images by Coupling Superpixel Context Information with Artificial Neural Networks	317
<i>Ahmad A. Hammoudi, Fuhai Li, Liang Gao, Zhiyong Wang, Michael J. Thrall, Yehia Massoud, and Stephen T.C. Wong</i>	
3D Segmentation in CT Imagery with Conditional Random Fields and Histograms of Oriented Gradients	326
<i>Chetan Bhole, Nicholas Morsillo, and Christopher Pal</i>	
Automatic Human Knee Cartilage Segmentation from Multi-contrast MR Images Using Extreme Learning Machines and Discriminative Random Fields	335
<i>Kunlei Zhang and Wenmiao Lu</i>	
MultiCost: Multi-stage Cost-sensitive Classification of Alzheimer's Disease	344
<i>Daoqiang Zhang and Dinggang Shen</i>	

Classifying Small Lesions on Breast MRI through Dynamic Enhancement Pattern Characterization	352
<i>Mahesh B. Nagarajan, Markus B. Huber, Thomas Schlossbauer, Gerda Leinsinger, Andrzej Krol, and Axel Wismüller</i>	
Computer-Aided Detection of Polyps in CT Colonography with Pixel-Based Machine Learning Techniques.....	360
<i>Jian-Wu Xu and Kenji Suzuki</i>	
Author Index	369

Learning Statistical Correlation of Prostate Deformations for Fast Registration

Yonghong Shi^{1,2}, Shu Liao¹, and Dinggang Shen^{1,*}

¹ IDEA Lab, Department of Radiology and BRIC, University of North Carolina at Chapel Hill
dgshen@med.unc.edu

² Digital Medical Research Center, Shanghai Key Lab of MICCAI, Fudan University, China

Abstract. This paper presents a novel fast registration method for aligning the planning image onto each treatment image of a patient for adaptive radiation therapy of the prostate cancer. Specifically, an online correspondence interpolation method is presented to learn the statistical correlation of the deformations between prostate boundary and non-boundary regions from a population of training patients, as well as from the online-collected treatment images of the same patient. With this learned statistical correlation, the estimated boundary deformations can be used to rapidly predict regional deformations between prostates in the planning and treatment images. In particular, the population-based correlation can be initially used to interpolate the dense correspondences when the number of available treatment images from the current patient is small. With the acquisition of more treatment images from the current patient, the patient-specific information gradually plays a more important role to reflect the prostate shape changes of the current patient during the treatment. Eventually, only the patient-specific correlation is used to guide the regional correspondence prediction, once a sufficient number of treatment images have been acquired and segmented from the current patient. Experimental results show that the proposed method can achieve much faster registration speed yet with comparable registration accuracy compared with the thin plate spline (TPS) based interpolation approach.

Keywords: Adaptive radiation therapy, Fast registration, Patient-specific statistical correlation, Canonical correlation analysis.

1 Introduction

Prostate cancer is the second-leading cause of cancer death for American men [1]. External beam radiation therapy is often used for prostate cancer treatment [2] by spreading the treatment over a weeks-long series of daily fractions. Since the prostate is surrounded by healthy tissue that can also be harmed by radiation, it is important to maximize the harm to the cancer while minimize the harm to healthy tissue. In addition, for adaptive radiation therapy, it is important to adjust the treatment plan by updating beam intensities and shapes according to the estimated patient position and motion from the acquired images [3, 4]. All of these tasks need image segmentation and registration for prostate, as studied extensively in the literature.

* Corresponding author.

In particular, image registration is able to estimate prostate deformation during radiotherapy. Thus with registration between the planning image and each treatment image, the treatment plan determined in the planning image space can be warped onto the treatment image space, and also the dose distribution map in the treatment image space can be warped back onto the planning image space for measurement of the actual total dose delivered to the patient after a portion of the treatment [5-7].

To achieve this goal, the segmented prostate in the planning image should be fast registered onto the prostate in each treatment day or vice versa. However, there are two difficulties with current registration algorithms. First, due to the limited contrast of CT images around the prostate boundaries, it is very difficult for the intensity-based registration algorithms to differentiate the corresponding structures in the planning and treatment images [5, 6]. Second, it is also difficult to perform fast registration by using the conventional intensity-based registration algorithms [7, 8], or even using the interpolation algorithms such as thin plate spline (TPS) [9] when the number of correspondences in the prostate boundaries or the size of region of interest (ROI) become large.

Therefore, in this paper, we present a novel fast registration method, which can learn the statistical correlation of the deformations between prostate boundary and non-boundary regions from the training patients, as well as from the online-collected treatment images of the current patient. With the learned statistical correlation, the estimated boundary correspondences can be used to rapidly predict the non-boundary regional correspondences between prostates in the planning and treatment images. Also, the patient-specific correlation information can be updated online to reflect the prostate shape changes during the treatment. The proposed method is detailed below.

2 Method

2.1 Description

Statistical shape-based segmentation algorithm [10] can *not only* segment the prostate from the planning and treatment images, *but also* obtain boundary correspondences between all segmented prostate shapes. To establish dense correspondences (or transforms) for the points around the prostate shapes, the use of correspondence interpolation algorithms such as TPS is needed. However, these algorithms are generally slow and not efficient for clinical application.

Actually, the statistical correlation between the boundary correspondences and non-boundary regional correspondences can be learned from a population of training patients, as well as available treatment images of the current patient. Then, this learned correlation can be used to guide the estimation of dense correspondences for the non-boundary regions around the prostate. In the following, we will detail this idea. In particular, Section 2.2 will first introduce how to learn the statistical correlation from a population of training patients. Then, Section 2.3 will introduce how to learn the patient-specific correlation from a patient, and also how to use the patient-specific correlation to gradually replace the population information for more effective interpolation of dense correspondences for prostate.

2.2 Learning Statistical Correlation from a Population of Training Patients

Training Samples: Given a number of prostate samples segmented from serial images of training patients, the boundary correspondences across different prostate shapes can be established by the deformable shape model [10]. Similarly, we can estimate the correspondences for the prostate shape segmented from the planning image of the current patient. In this way, we can use the established correspondences to warp the prostate shapes of each training patient onto the planning image space of the current patient, for obtaining the aligned prostate shapes.

Let us assume that we have P training patients $\{S^i|i = 1, \dots, P\}$. Each patient has d_i segmented prostate images $\{S_j^i|i = 1, \dots, P; j = 1, \dots, d_i\}$, where each image is represented by a shape with M points, i.e., $S_j^i = \{(x_{j,l}^i, y_{j,l}^i, z_{j,l}^i)|l = 1, \dots, M\}$. We can linearly align each prostate shape S_j^i of the i -th training patient onto the prostate shape S_1^0 of the planning image of the current patient S^0 , for obtaining its aligned prostate shape $S_{j \rightarrow 1}^i$ in the planning image space of the current patient. Based on all aligned prostate shapes $\{S_{j \rightarrow 1}^i\}$ from each patient S^i , we can calculate their mean shape $\bar{S}^i = \sum_{j=1}^{d_i} S_{j \rightarrow 1}^i / d_i$, and then obtain the residual deformation $B_j^i = S_{j \rightarrow 1}^i - \bar{S}^i$ from each aligned prostate shape $S_{j \rightarrow 1}^i$ by subtracting the mean prostate \bar{S}^i from $S_{j \rightarrow 1}^i$. Thus, based on the prostate shape in the planning image S_1^0 of the current patient S^0 and the learned residuals from different samples of training patients, we can obtain various deformed prostate shapes, $\{C_j^i = B_j^i + S_1^0|i = 1, \dots, P; j = 1, \dots, d_i\}$, to simulate possible prostate deformations for the current patient during the treatment.

Then, by using a correspondence interpolation algorithm, i.e., TPS [11], we can interpolate N dense regional correspondences for each deformed prostate shape C_j^i . Accordingly, we can get many paired samples for boundary correspondences and non-boundary regional correspondences, i.e., $\{(C_j^i, R_j^i)|i = 1, \dots, P; j = 1, \dots, d_i\}$, where C_j^i is the $3M$ -dimensional vector and R_j^i is the $3N$ -dimensional vector. With this training set, we can first perform principal component analysis (PCA) to get a compact representation for the boundary correspondences by using all samples $\{C_j^i|i = 1, \dots, P; j = 1, \dots, d_i\}$, as well as for the non-boundary correspondences using the samples $\{R_j^i|i = 1, \dots, P; j = 1, \dots, d_i\}$ as described in Equations (1) and (2) below.

$$C_j^i = \bar{C} + \varphi_C \alpha_j^i \quad (1)$$

$$R_j^i = \bar{R} + \varphi_R \beta_j^i \quad (2)$$

Where \bar{C} and \bar{R} denote the respective means for $\{C_j^i\}$ and $\{R_j^i\}$, with φ_C and φ_R as the respective eigenvector matrices. α_j^i and β_j^i denote the respective reconstruction coefficients vectors for $\{C_j^i\}$ and $\{R_j^i\}$. Thus, each prostate shape C_j^i can be represented by a small number of coefficients, i.e., a coefficient vector α_j^i , and each non-boundary regional correspondence sample can also be represented by a

coefficient vector β_j^i . We can then estimate the correlation between the coefficient vectors α_j^i and β_j^i by using canonical correlation analysis (CCA) [11].

CCA-based Prediction: CCA explores the correlative structures of two sets of variables, i.e., α_j^i and β_j^i in our application. The canonical correlation is the correlation of two canonical (latent) variables, one variable representing a set of independent variables, and the other representing a set of dependent variables. The canonical correlation is optimized such that the linear correlation between the two latent variables is maximized. Mathematically, CCA extracts the correlated modes between vectors α_j^i and β_j^i by seeking a set of transformation vector pairs, A_k and B_k , which yields the canonical variates u_k and v_k with maximum correlation as follows:

$$u_k = (\alpha_j^i)^T A_k \quad (3)$$

$$v_k = (\beta_j^i)^T B_k \quad (4)$$

With estimation of canonical variates u_k and v_k , we can estimate their correlation and use it for prediction. For instance, given a new prostate shape segmented from a new treatment image of the current patient at t -th time-point S_{t-1}^0 , we can first obtain its PCA-based coefficient vector α_t^0 according to Equation (1), and then obtain the corresponding canonical variates u_k using Equation (3). The estimation of the canonical variates v_k is in the canonical space. The linear regression is used for the estimation of v_k from u_k [11]. Then, we can estimate the coefficient vector β_t^0 for the non-boundary regional correspondences using Equation (4), and use the reversed PCA to eventually estimate the regional correspondences for the prostate in the new treatment image R_t^0 by Equation (2).

2.3 Refining Interpolation of Dense Correspondences by Patient-Specific Information

For each treatment image of the current patient, we can first segment the prostate shape, and then establish the correspondences between the prostate in the planning image and the prostate in the treatment image. The new pairs of correspondences in the prostate boundaries and the non-boundary region around the prostate can be used to train a patient-specific correlation model between the deformations of the prostate boundaries and non-boundary regions. Since this model is more patient specific, its contribution should become larger than that from the population of training patients, once enough treatment images have been collected from the current patient.

Assume that we have $t - 1$ segmented prostate images $\{S_j^0 | j = 1, \dots, t - 1\}$ from the current patient S^0 . Then, we can obtain their aligned prostate shapes $\{S_{j-1}^0 | j = 1, \dots, t - 1\}$ in the planning image (S_1^0) space, the mean prostate shape $\bar{S}^0 = \sum_{j=1}^{t-1} S_{j-1}^0 / (t - 1)$, the residual deformation $B_j^0 = S_{j-1}^0 - \bar{S}^0$, and eventually all possible patient-specific prostate deformations for the current patient, i.e., $\{C_j^0 = B_j^0 + S_1^0 | j = 1, \dots, t - 1\}$. Similarly, the respective non-boundary regional correspondences $\{R_j^0 | j = 1, \dots, t - 1\}$ can also be obtained with TPS. In this way, we

can obtain a set of patient-specific samples, $\{(C_j^0, R_j^0) | j = 1, \dots, t-1\}$, to train a patient-specific correlation model, as done for the training patients in Section 2.2.

Accordingly, our strategy for collecting the statistical correlation and predicting the dense correspondences R_t for the current patient at the t -th time point can be described in Equation (5). We can initially rely more on the population-based correlation model, and then gradually reduce its contribution with the collection of more and more patient-specific information.

$$R_t = (1 - \omega_t)R_t^0 + \omega_t(R_t^0)' \quad (5)$$

$$\omega_t = \begin{cases} 0 & t \leq N_s \\ (t - N_s)/(N_b - N_s) & N_s \leq t \leq N_b \\ 1 & N_b < t \end{cases} \quad (6)$$

Here, R_t^0 denotes the non-boundary regional correspondences predicted by the population-based correlation model as described in Section 2.2, while $(R_t^0)'$ denotes the correspondences predicted by the patient-specific correlation model learned from $t-1$ treatment images of the current patient. The term $(R_t^0)'$ is not existent for several initial treatment images, simply because the available segmented prostate images for the specific patient are not sufficient to train a statistical correlation model. We can actually begin to train the patient-specific statistical correlation model once N_s time-point images have been obtained. Here, N_s is the minimal number of treatment images for the current patient, and $N_s = 3$ is used in our study. Afterward, we can gradually increase the weight of term $(R_t^0)'$ and simultaneously decrease the weight of term R_t^0 , as more and more new treatment images of the current patient are collected. This can be achieved by defining parameter ω_t according to Equation (6). If we have more than N_b treatment images from the same patient, we can stop using the population-based estimation term R_t^0 , since the patient-specific correlation collected from over N_b treatment images is enough to predict the non-boundary regional correspondence.

Our method for fast correspondence interpolation is summarized as follows:

- 1) Use the deformable segmentation method to segment prostate from the planning image of a patient under treatment.
- 2) Linearly align all prostate shape samples of training patients onto the planning image space of the current patient, and then use the aligned samples to learn a population-based correlation model, for capturing the relation between the boundary correspondences and the non-boundary regional correspondences of prostate.
- 3) For each new treatment image acquired from the current patient, we will first segment the prostate and establish its boundary correspondences with that in the planning image. Then, with the population-based correlation model as well as the established correspondences on the prostate boundary, we can estimate the non-boundary regional correspondences in the prostate of the current patient.
- 4) With segmentation of more treatment images, we can use them to learn a patient-specific correlation model, to work with the population-based correlation model for correspondence prediction. Once a sufficient number of the treatment images are segmented, the patient-specific correlation model can take over the task for correspondence prediction, without using the population-based correlation model.

3 Experimental Results

The performance of our fast online correspondence interpolation method is tested on serial prostate CT images of 24 patients. Most patients have up to 12 scan images, each with image size of $512 \times 512 \times 61$ and voxel size of $1 \times 1 \times 3\text{mm}^3$. Prostates in all serial images of all patients have been manually segmented by medical expert, and their correspondences have been established with a statistical shape-based method. For comparison, TPS-based correspondence interpolation is used as a baseline method [9, 13]. Both qualitative and quantitative results are reported.

Results on an Individual Patient: Fig. 1 shows the predictive error of prostate deformations in the serial images of a patient by our method, compared to those interpolated by TPS. The top row shows the predictive errors at different treatment times using only the population-based statistical correlation model by setting $\omega_t = 0$ in Equation (5). The bottom row shows the predictive errors using both population-based and patient-specific statistical correlation models. In this figure, the pink contours represent prostate boundaries, while the light-blue outer contour and the white interior contour represent the iso-distance of 15mm from the prostate boundary contour, respectively. According to the color bar in Fig. 1, our fast prediction method can achieve a prediction error of less than 1.0mm.

Fig. 2 shows quantitatively the error distributions for results in Fig. 1. It can be observed that the error distributions in all treatment images are similar to each other when using only the population-based correlation model (red bar). In contrast, the prediction errors are decreased with adding of more and more treatment images of the current patient (blue bar).

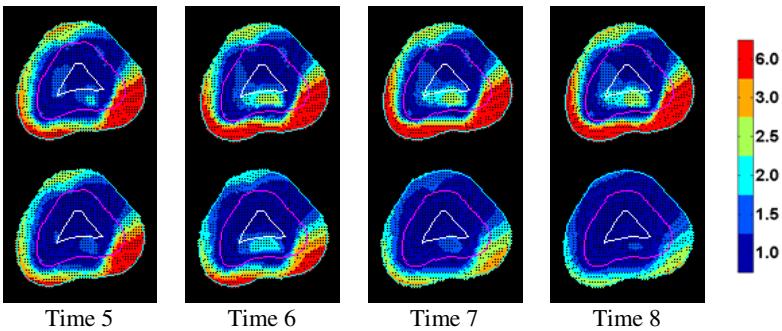


Fig. 1. The distribution of predictive errors at different treatment times by using only the population-based (top row) or both population-based and patient-specific statistical correlation (bottom row). Pink contour represents the shape of prostate boundary, while the light blue and white contours represent the iso-distance of 15mm from the prostate boundary contour.

Results on All 24 Patients: Table 1 shows the overall predictive errors for all images of all 24 patients by using only the population data, or both population and patient data. The predictive errors on the voxels around 5mm of prostate boundary are calculated, since these voxels locate between prostate and normal tissues where accurate registration is critical for effective treatment, i.e., maximizing the harm to the cancer and minimizing the harm to healthy tissue. It can be observed that the average

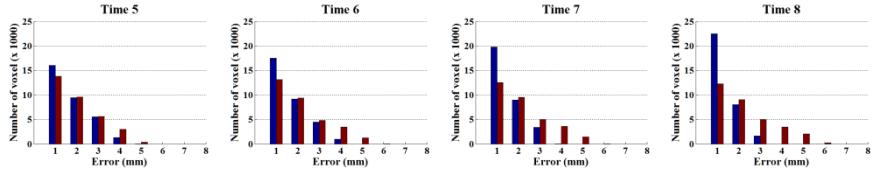


Fig. 2. Overall distributions of predictive errors at 4 different treatment images of a patient, by using only the population-based (red) or both population-based and patient-specific statistical correlation (blue)

error is 0.38mm by our method using both population-based and patient-specific statistical correlation, which is much better than 1.02mm obtained by using only the population-based statistical correlation. Also, the maximum error can be reduced significantly by our method.

Table 2 shows the effect on the predictive error with respect to the number of patient's treatment images used for estimating deformation correlation. As we can observe, the predictive error of prostate deformation is reduced with use of more treatment images for learning the patient-specific information.

Table 1. Comparison of predictive errors on all images of all 24 patients between the method using only population data and the method using both population and patient data

	Mean \pm Std	Minimum	Median	Maximum
Population	1.02 \pm 0.69	0.003	0.88	5.02
Population + Patient-specific	0.38 \pm 0.27	0.001	0.27	3.14

Table 2. The effect on the predictive error with respect to the number of patient's treatment images used for estimation of deformation correlation

Number of patient's treatment images used	Mean \pm Std	Minimum	Median	Maximum
3	0.62 \pm 0.42	0.005	0.53	3.14
4	0.55 \pm 0.37	0.004	0.48	2.37
5	0.45 \pm 0.30	0.002	0.39	1.85
6	0.32 \pm 0.23	0.003	0.27	1.35
7	0.30 \pm 0.24	0.007	0.24	1.72
8	0.27 \pm 0.20	0.007	0.21	1.89
9	0.22 \pm 0.22	0.001	0.17	2.50

Speed: The above results demonstrate that our online correspondence interpolation method performs comparable to TPS. For comparing the speed, we load the same prostate boundary landmarks and correspondences into our method and TPS [13] to estimate the dense deformations. As shown in Table 3, when considering larger ROI with the size of 512x512x61, our method is able to predict the dense deformation field within 24.50 seconds, while TPS needs 6.70 minutes; when considering small

ROI with the size of $112 \times 110 \times 93$ exactly around the prostate, it takes 1.85 seconds for our method to estimate the deformation field while TPS needs 25.0 seconds. This result indicates that our method can estimate the dense deformation field in real-time, which has potential to be applied to clinical applications. (OS: 32 bit of Windows 7, CPU: Intel Core 2 Quad Q9400 2.66Hz, Memory: 4GB).

Table 3. Comparison the running times for our method and TPS

Prostate data (Number of Landmarks: 816)	TPS	Our method
Large ROI (Size: $512 \times 512 \times 61$)	6.70 minutes	24.50 seconds
Small ROI (Size: $112 \times 110 \times 93$)	25.00 seconds	1.85 seconds

4 Conclusion

We have presented a new fast registration for aligning prostates in the planning image and each treatment image of a patient during radiotherapy, by learning the statistical correlation of the deformations between prostate boundaries and non-boundary regions from both population and patient data. The patient-specific statistical correlation is online-and-incrementally learned from the previous treatment image of the current patient. For the initial treatment images, the population-based statistical correlation plays major function for statistically predicting the dense regional correspondence. As more and more segmented prostate images are obtained from the current patient, the patient-specific statistical correlation starts to take an important role for predicting the dense regional correspondences. Experimental results on real patient data show that our method can produce comparable accuracy as TPS, but it performs much faster than TPS. In the future, we will incorporate our method into the whole pipeline for prostate segmentation and registration, and will further test its overall performance and speed as well.

Acknowledgement. This research was supported by the grant from National Institute of Health (Grant No. 1R01 CA140413). The image data and expert contours used in this study were provided by the Department of Radiation Oncology, UNC-Chapel Hill under support of NIH grants R01 RR018615 and R44/43 CA119571, E.L. Chaney PI. This work was also supported by the grant from National Natural Science Foundation of China (No. 60972102).

References

1. Prostate Cancer Foundation, <http://www.prostatecancerfoundation.org/>
2. Perez, C.A., Brady, L.W., Halperin, E.C., Schmidt-Ullrich, R.K.: Principles and Practice of Radiation Oncology, 4th edn. Lippincott Williams & Wilkins (2004)
3. Yan, D., Lockman, D., Brabbins, D., Tyburski, L., Martinez, A.: An Off-line Strategy for Constructing a Patient-specific Planning Target Volume in Adaptive Treatment Process for Prostate Cancer. Int. J. Radiat. Oncol. Biol. Phys. 48, 289–302 (2000)

4. Court, L.E., Tishler, R.B., Petit, J., Cormack, R., Chin, L.: Automatic Online Adaptive Radiation Therapy Techniques for Targets with Significant Shape Change: a Feasibility Study. *Phys. Med. Biol.* 51, 2493–2501 (2006)
5. Yan, D., Jaffray, D.A., Wong, J.W.: A Model to Accumulate Fractionated Dose in a Deforming Organ. *Int. J. Radiat. Oncol. Biol. Phys.* 44, 665–675 (1999)
6. Schaly, B., Kempe, J.A., Bauman, G.S., Battista, J.J., Dyk, J.V.: Tracking the Dose Distribution in Radiation Therapy by Accounting for Variable Anatomy. *Phys. Med. Biol.* 49, 791–805 (2004)
7. Foskey, M., Davis, B., Goyal, L., Chang, S., Chaney, E., Strehl, N., Tomei, S., Rosenman, J., Joshi, S.: Large Deformation Three-dimensional Image Registration in Image-Guided Radiation Therapy. *Phys. Med. Biol.* 50, 5869–5892 (2005)
8. Christensen, G.E., Rabbitt, R.D., Miller, M.I.: Deformable Templates Using Large Deformation Kinematics. *IEEE Trans. Image Processing* 5, 1435–1447 (1996)
9. Bookstein, F.L.: Principal Warps: Thin-Plate Splines and the Decomposition of Deformations. *IEEE Trans. Pattern anal. Mach. Intell.* 11, 567–585 (1989)
10. Feng, Q., Foskey, M., Chen, W., Shen, D.: Segmenting CT Prostate Images Using Population and Patient-specific Statistics for Radiotherapy. *Med. Phys.* 37(8), 4121–4132 (2010)
11. Hotelling, H.: Relations between Two Sets of Variates. *Biometrika* 28, 321–377 (1936)
12. Liu, T., Shen, D., Davatzikos, C.: Predictive Modeling of Anatomic Structures Using Canonical Correlation Analysis. In: Proceedings of the IEEE International Symposium on Biomedical Imaging, VA, Arlington, pp. 1279–1282 (2004)
13. ITK Insight Toolkit, <http://www.itk.org/>

Automatic Segmentation of Vertebrae from Radiographs: A Sample-Driven Active Shape Model Approach

Peter Mysling¹, Kersten Petersen¹, Mads Nielsen^{1,2}, and Martin Lillholm²

¹ DIKU, University of Copenhagen, Denmark

mysling@diku.dk

² BiomedIQ, Rødovre, Denmark

Abstract. Segmentation of vertebral contours is an essential task in the design of automatic tools for vertebral fracture assessment. In this paper, we propose a novel segmentation technique which does not require operator interaction. The proposed technique solves the segmentation problem in a hierarchical manner. In a first phase, a coarse estimate of the overall spine alignment and the vertebra locations is computed using a shape model sampling scheme. These samples are used to initialize a second phase of active shape model search, under a nonlinear model of vertebra appearance. The search is constrained by a conditional shape model, based on the variability of the coarse spine location estimates. The technique is evaluated on a data set of manually annotated lumbar radiographs. The results compare favorably to the previous work in automatic vertebra segmentation, in terms of both segmentation accuracy and failure rate.

1 Introduction

Prevalent and incident vertebral fractures are of clinical importance in osteoporosis pharmaceutical trials. In current practise, vertebral fracture assessment is often performed manually using quantitative morphometry from lateral radiographs. This process is both labor-intensive and prone to false positives, as a result of relying on a 6-point representation of each vertebra. Automatic segmentation of the vertebral contours could be an essential step in reducing the false-positive rates and in fully automating the fracture assessment process.

Previously, several accurate, semi-automatic techniques based on active shape models (ASMs) [5] and active appearance models (AAMs) [4] have been proposed [9,12,13]. These techniques do, however, require manual initialization of between 1 and 6 points for each vertebra. Fully automatic procedures, which operate at lower accuracy, have been explored in [6,14].

In this paper, we propose a hierarchical, fully automatic technique for segmentation of the vertebral outlines in lateral radiographs. The proposed technique makes use of machine learning methods both in learning generative models of the vertebral shape variation and in learning classification-based models of the

vertebral appearance. The performance of the technique compares favorably to the previous work in automatic vertebral segmentation.

2 Methods

The proposed technique models the spine shape and appearance on two separate levels. The overall spine alignment and vertebra locations are estimated using a sampling scheme, from a low-resolution spine model based on 6 points for each vertebra. A detailed segmentation of the corresponding vertebral contours is determined from a high-resolution model of individual vertebrae using an ASM approach. The ASM-search is efficiently constrained by a conditional model of valid shape variation, given a coarse vertebra location estimate. Sample manual annotations of both the low- and high-resolutinal models are given in Figure 1(a). Each component of the proposed technique is described in the following.

2.1 Statistical Shape Models

We employ traditional linear point distribution models (PDMs), as originally proposed by Cootes et al. in [5]. Consider a training set of N shapes, where each shape, \mathbf{x} , is parametrized by a sequence of n landmarks, such that $\mathbf{x} = (x_1, y_1, \dots, x_n, y_n)^T$. To establish a common frame of reference in which all training shapes are aligned, each training shape undertakes a translation (t_x, t_y) , a scaling s , and a rotation r . The optimal alignment transformations are determined by generalized Procrustes analysis (GPA) [8]. We refer to these transformations as the *pose parameters* and parametrize them as $\mathbf{a} = (t_x, t_y, s, r)^T$. The aligned shapes are summarized in a *reference shape*. This shape facilitates transformations between the image space and the space of normalized shapes.

A compact shape representation, which retains most of the shape variation in the training set, is computed by principal component analysis (PCA). Accordingly, the *shape parameters*, \mathbf{b} , which account for pose-free shape variation are given by $\mathbf{b} = \mathbf{P}^T(\mathbf{x} - \bar{\mathbf{x}})$, where $\bar{\mathbf{x}}$ denotes the mean shape and \mathbf{P} denotes the matrix of shape covariance eigenvectors. The linear transformation is chosen such that 99% of the shape variation is explained. We let $\theta = (\mathbf{a}^T, \mathbf{b}^T)^T \in \Theta$ denote the vector of concatenated model parameters.

2.2 Coarse Sample-Based Segmentation

We propose to model an initial estimate of the spine location by a collection of shape samples. In accordance with the hierarchical approach, the overall spine shape is modelled by a low-resolutinal spine PDM, based on six points for all modelled vertebrae. Visualization of a sample manual annotation of the L1 through L4 vertebrae is given in Fig. 1(b).

Following the work of Petersen et al. in [11], we approach the sampling problem using a statistical formulation of the shape fit to the image evidence and employ a static Sequential Monte Carlo (SMC) sampler [10] in sampling this distribution. This approach is similar to that of shape-based particle filtering

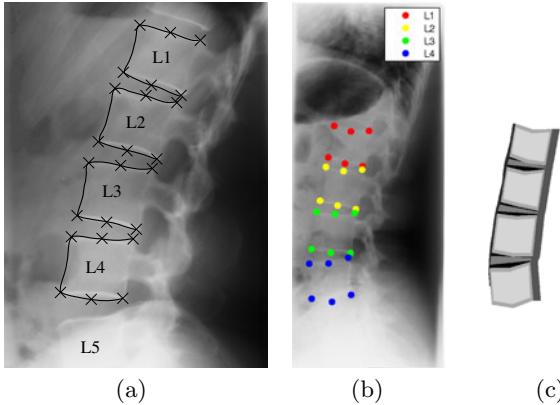


Fig. 1. Visualization of the manual annotations and template label assignment scheme. (a) Manual six-point and full boundary annotations. (b) Manual six-point annotations. (c) Shape template, as implied by the manual annotations of (b). The spine is divided into 6 classes, as indicated by the region intensities. The classes are anterior, posterior, and overall background, intra- and inter-vertebral space, and the vertebral boundary.

segmentation—as originally proposed in [6]—albeit with two improvements: 1) the SMC-sampler on shapes applies a statistically motivated formulation of the importance weights and 2) the SMC sampler exploits the static nature of images.

The SMC sampler on shapes relies on a pixel classification-based model of the appearance inside, and surrounding, the object of interest. Class labels are assigned to each pixel according to their position relative to the spine shape $\theta_s \in \Theta_s$. We refer to such shape-guided label assignment schemes as *shape templates*. The shape template for the spine segmentation is visualized in Figure 1(c).

This region-based formulation of the object appearance allows us to evaluate the global consistency of candidate segmentations. A pixel classifier that distinguishes between the regions of interest is trained from the input data $\mathcal{D} = \{\mathbf{x}_n\}_{n=1}^N$ with associated labels $\mathbf{u} = (u_1, \dots, u_N) \in \mathcal{Y}^N$, where \mathcal{Y} denotes the space of class labels. The quality of a candidate shape segmentation $\theta_s^{cand} \in \Theta$ is summarized by multiplying the likelihood of the individual pixels, under the template labeling implied by the shape θ_s^{cand} . Following [11], we employ a random forest pixel classifier [3] and a feature representation based on Gaussian derivatives on multiple scales [7].

In [11], it is shown that, under the assumption of i.i.d. data \mathcal{D} and conditional independence of $p(\mathbf{u}|\theta_s)$, the shape parameter posterior can be formulated as

$$p(\theta_s|\mathcal{D}) \propto p(\theta_s) \prod_{n=1}^N \sum_{u_n \in \mathcal{Y}} p(\mathbf{x}_n|u_n, \theta_s) p(u_n|\theta_s) \equiv \pi(\theta_s|\mathcal{D}).$$

The likelihood term $p(\mathbf{x}_n|u_n, \theta_s)$ is computed from the trained pixel classifier. The prior term $p(\theta_s)$ is estimated from the shape training data under the assumption of normally distributed pose and shape. The $p(u_n|\theta_s)$ term is simply

computed from the shape template associated to θ_s and can be thought of as a class-indicator.

The static SMC sampler efficiently samples from a sequence of distributions $\{\pi_t\}_{t=1}^T$. This sequence is designed as a bridge between the proposal distribution π_1 and the target distribution $\pi_T = \pi$ through a sequence of intermediate artificial distributions. These artificial distributions are constructed using an annealing scheme, such that $\pi_t(\theta_s) = \pi(\theta_s | \mathcal{D})^{\beta_t} p(\theta_s)^{1-\beta_t}$, where the annealing parameters $\{\beta_1, \dots, \beta_T\}$ must satisfy $0 \leq \beta_1 < \dots < \beta_T = 1$. In this manner, samples are gradually shifted towards regions of high density.

During each iteration, the SMC sampler maintains a set of M samples, $\{\theta_{s_t}^{(l)}\}_{l=1}^M$, with associated importance weights $\{w_t^{(l)}\}_{l=1}^M$ which approximately summarizes π_t . Initially, the sample set $\{\theta_{s_1}^{(l)}\}_{l=1}^M$ is simply generated by sampling the shape and pose prior $p(\theta_s)$ and weighted according to $\pi_1(\theta_s)$. During subsequent iterations, all weights $\{w_t^{(l)}\}_{l=1}^M$ are first computed according to

$$\frac{w_t}{w_{t-1}} \propto \frac{\pi_t(\theta_{s_{t-1}})}{\pi_{t-1}(\theta_{s_{t-1}})}.$$

Following this, the sample set $\{\theta_{s_t}^{(l)}\}_{l=1}^M$ is resampled according to the normalized weights. Duplicate samples are redrawn from a PDM of the local shape variation in the neighborhood of that sample.

When this procedure terminates, the sample set $\{\theta_{s_T}^{(l)}\}_{l=1}^M$ approximates the model posterior $\pi(\theta_s | \mathcal{D})$, under the weighting of $\{w_{s_T}^{(l)}\}_{l=1}^M$. The complete sample set can either be retained for further computations or, alternatively, be summarized as the MAP estimate by retaining the sample of maximum weight.

2.3 Conditional Shape Model

We wish to construct a statistical model of the most likely vertebra contour variation, given a coarse estimate of the corners and end-plate mid-points of the individual vertebrae. This model allows us to initialize and constrain the following ASM-search in an efficient manner. Previous work on conditional vertebra models indicate that the variation is efficiently modelled as a conditional normal distribution on the shape parameters [9]. We follow the same approach.

The statistical model relies on the parametrization of two PDMs. The low-dimensional PDM is simply based on the individual vertebrae of the coarse segmentations. The high-resolitional PDM is based on 66 equidistant points delineating the vertebrae. Sample manual annotations are given in Fig. 1(a).

One effect of modelling the shape likelihood as a conditional model on the shape parameters is that the conditional model is pose free. Consequently, pose-correspondence must be established between the two PDMs, such that the pose parameters of the low-dimensional model can be directly applied in the high-dimensional model. We propose to achieve pose correspondence using the following two-step approach:

1. Two intermediate high-resolutonal reference shapes are computed: 1) a pose-corresponding reference shape, \mathbf{x}_H^{PC} , where the pose transformations have been inherited from the low-resolutonal model and 2) an independent reference shape, \mathbf{x}_H^I , based on regular alignment of all landmarks in the high-dimensional training set.
2. The final reference shape, \mathbf{x}_H , is computed by aligning the independent reference shape, \mathbf{x}_H^I , to the pose-corresponding reference shape, \mathbf{x}_H^{PC} .

By constructing the high-dimensional PDM from the resulting reference shape, we achieve both pose-correspondence between the high- and low-dimensional models and landmark correspondence in the high-dimensional model.

Having established the required shape models, we now consider the conditional model on the shape parameters. Let $\theta_L = [\mathbf{a}_L^T, \mathbf{b}_L^T]^T$ and $\theta_H = [\mathbf{a}_H^T, \mathbf{b}_H^T]^T$ denote the low- and high-resolutonal vertebra representations, respectively. We model $p(\mathbf{b}_H|\mathbf{b}_L)$, the conditional distribution of the shape parameters, as the Gaussian density $p(\mathbf{b}_H|\mathbf{b}_L) = \mathcal{N}(\boldsymbol{\mu}_{cond}, \mathbf{C}_{cond})$. The conditional mean and conditional covariance can be computed using standard matrix techniques.

2.4 Active Shape Model Segmentation

We compute a detailed segmentation of the individual vertebrae by a multi-resolution active shape model (ASM) [5]. The ASM-search is carried out in the constrained shape space given by the conditional shape model. This ASM model has previously been described by Iglesias et al. in [9]. To accommodate for the inaccuracy of the coarse segmentations of the SMC-sampler, we introduce three modifications to the model of Iglesias et al.

Firstly, we retain the complete sample set $\{\theta_{st}\}_{l=1}^M$ from the SMC sampler on shapes and initialize M separate ASM-searches for each vertebra. Each search is initialized by sampling the low-resolutonal shapes $\{\theta_{st}\}_{l=1}^M$ according to their weights $\{w_T\}_{l=1}^M$ and applying the conditional shape model. We choose the final model as the segmentation of maximum appearance fit. Secondly, the ASM landmark updates are regularized according to the dynamic programming formulation of Behiels et al. [2]. Thirdly, we employ a nonlinear appearance model based on a random forest classifier. This choice was taken on account of the highly nonlinear appearance along the vertebral boundaries. The adapted ASM procedure is outlined in the following.

For each landmark position, we train a random forest classifier that distinguishes between profiles from two classes; boundary and nonboundary. Examples of both classes are sampled perpendicular to the vertebral contours. During segmentation, the ASM procedure computes the posterior probability for a range of viable landmark displacements. Letting $d_i \in \{-n_s, \dots, n_s\}$ denote the displacement of the i th landmark, the displacement configurations are chosen under a regularized expression of the appearance fit, such that

$$\underset{d_1, \dots, d_n}{\operatorname{argmax}} p_1(b|d_1) + \sum_{i=2}^n (p_i(b|d_i) - \alpha|d_i - d_{i-1}|).$$

Here, $p_i(b|d_i)$ denotes the posterior probability of the displacement d_i belonging to the boundary class under the random forest model of the i th landmark and α is a regularization parameter which controls the outlier punishment. Optimal settings of d_1, \dots, d_n is efficiently computable by dynamic programming [2].

To ensure that only plausible shapes are generated, segmentations are constrained by the Mahalanobis distance induced by the conditional shape model.

3 Evaluation

The proposed technique was evaluated on a data set of 157 digitized lateral radiographs of the lumbar spine. The acquired radiographs are a subset from the PERF cohort [1]. Manual 6-point and full contour delineation annotations for the L1 to L4 vertebrae were marked by trained radiologists.

We compute an unbiased estimate of the technique performance by training and testing our segmentation model in a nested 6-fold cross-validation fashion. Each fold of the cross-validation scheme includes an independent round of greedy model parameter tuning from a range of viable parameters.

The error of the resulting segmentations is reported as the average point-to-contour distance, i.e., the average distance between the landmarks of the segmentation to the closest point on the polygonal line of the manual annotations. This is the current standard in vertebra segmentation evaluation [6,9,12,13,14].

The vertebrae of the spinal column are similar in appearance and differ mostly in size. Consequently, the procedure will occasionally displace the segmentation one level up or down, e.g., the L2 vertebra will be located as L1 and so forth. We refer to this phenomenon as *level-shifting*. Following the previous work in automatic vertebra segmentation [6,14], the segmentations are divided into 3 classes: successful, level-shifted and failures. For successful segmentations, the vertebra center must be located within 10 mm of the true center. Level-shifted segmentations are those that have been shifted up or down one or more levels along the vertebra column, but are otherwise successful. All remaining segmentations are treated as failures. Segmentation accuracy is only evaluated for successful cases.

A comparison of the presented technique to the previous work in vertebra segmentation is given in Table 1. The techniques are categorized as fully automatic if they require no operator interaction and semi-automatic if the search procedure is initialized from manual indications of the vertebra locations. Numbers that were not reported in the original papers are indicated as such.

On the presented data set we achieve an average point-to-contour error of 0.73 mm for normal vertebrae and 2.17 mm for fractured vertebrae, resulting in a total average error of 0.81 mm. 46% of the produced segmentations have been level-shifted. The technique does, however, not produce any search failures.

4 Discussion and Conclusions

The results in Table 1 show that the accuracy of the proposed technique compares favorably to the previous automatic vertebra segmentation techniques—in

Table 1. Comparison of the proposed method and the previous work in vertebra segmentation from radiographs. The errors are reported as mean point-to-contour errors.

Authors	Modality	Failure Normal Fractured Total			
		(%)	(%)	(mm)	(mm)
<i>Semi-automatic Techniques</i>					
Roberts et al. [13]	DXA	0%	0%	0.75	1.24
Roberts et al. [12]	X-ray	0%	0%	0.64	1.06
Iglesias et al. [9]	X-ray	0%	0%	0.47	0.54
<i>Automatic Techniques</i>					
de Bruijne et al. [6]	X-ray	N/A	N/A	N/A	N/A
Roberts et al. [14]	X-ray	20%	7%	0.93	2.27
This study	X-ray	46%	0%	0.73	2.17
					0.81

terms of both normal and fractured vertebrae. Compared to the semi-automatic techniques, the proposed technique is less accurate; this is especially evident for fractured vertebrae. Note that the search procedure of Iglesias et al. [9] is initialized from six points on each vertebrae. The semi-automatic techniques of Roberts et al. in [13] and [12] require manual initialization of a single point inside each vertebra. We note that the experiments have been performed on different data sets for all studies and that the results should be interpreted as such.

Previously, the number of level-shifted segmentations and search failures have only been reported in [14] by Roberts et al. Compared to this work, our method is more sensitive to the ambiguity of the vertebra levels but significantly less prone to complete search failures. Note, however, that Roberts et al. have manually cropped the images below L4 to eliminate candidate segmentations that have been displaced one vertebra level down. It is reasonable to assume that the level-shifting issue cannot be resolved without explicit modelling of additional anatomical structures, such as the hip and tailbone.

The non-level-shifted segmentations contained a total of 319 normal vertebrae and 17 fractured vertebrae. Over this relatively small sample of fractured vertebrae, we obtain only modest segmentation accuracy, with an error above 2 mm for 52% of the fractures. Fractures are challenging for several reasons. First of all, fractured vertebrae tend to vary greatly, both in terms of the vertebra shape and the appearance along the end-plates. Furthermore, fractures only occur rarely, resulting in under-training of the shape and appearance models. We speculate that the accuracy can be improved by increasing the number of fracture training examples and enhancing the shape models for segmentation of fractured vertebrae, e.g., by ensuring that the fractured vertebrae have a large contribution to the vertebra models.

In summary, we have presented a novel, fully automatic technique for segmentation of the vertebral contours in radiographs. The presented technique was evaluated, in an unbiased manner, on data set of manually annotated

radiographs. The evaluation showed that the proposed technique compares favorably to the previous work in automatic vertebra segmentation in terms of the segmentation accuracy and failure rate. Further work is, however, required in order to reduce the level-shifting rate.

Acknowledgments. The authors gratefully acknowledge the funding from the Danish Research Foundation (Den Danske Forskningsfond) supporting this work. The authors thank Paola Pettersen for the annotations.

References

1. Bagger, Y.Z., Tankó, L.B., Alexandersen, P., Hansen, H.B., Qin, G., Christiansen, C.: The Long-term Predictive Value of Bone Mineral Density Measurements for Fracture Risk is Independent of the Site of Measurement and the Age at Diagnosis: Results from the Prospective Epidemiological Risk Factors Study. *Osteoporosis International* 17, 471–477 (2006)
2. Behiels, G., Vandermeulen, D., Maes, F., Suetens, P., Dewaele, P.: Active Shape Model-Based Segmentation of Digital X-ray Images. In: Taylor, C., Colchester, A. (eds.) *MICCAI 1999*. LNCS, vol. 1679, pp. 128–137. Springer, Heidelberg (1999)
3. Breiman, L.: Random Forests. *Machine Learning* 45, 5–32 (2001)
4. Cootes, T.F., Edwards, G.J., Taylor, C.J.: Active Appearance Models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23, 681–685 (2001)
5. Cootes, T.F., Taylor, C.J., Cooper, D.H., Graham, J.: Active shape models—their training and application. *Comp. Vision and Image Understanding* 61, 38–59 (1995)
6. de Bruijne, M., Nielsen, M.: Image Segmentation By Shape Particle Filtering. In: *ICPR 2004*, pp. 722–725. IEEE Computer Society Press, Los Alamitos (2004)
7. Florack, L., Ter Haar Romeny, B., Viergever, M., Koenderink, J.: The Gaussian Scale-space Paradigm and the Multiscale Local Jet. *International Journal of Computer Vision* 18, 61–75 (1996)
8. Gower, J.C.: Generalized Procrustes Analysis. *Psychometrika* 40, 33–51 (1975)
9. Iglesias, J.E., de Bruijne, M.: Semiautomatic Segmentation of Vertebrae in Lateral X-rays Using a Conditional Shape Model. *Acad. Rad.* 14, 1156–1165 (2007)
10. Johansen, A.M., Del Moral, P., Doucet, A.: Sequential Monte Carlo Samplers for Rare Events. Technical Report, Uni. of Cambridge, Dept. of Engineering (2005)
11. Petersen, K., Nielsen, M., Brandt, S.S.: A Static SMC Sampler on Shapes for the Automated Segmentation of Aortic Calcifications. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010*. LNCS, vol. 6314, pp. 666–679. Springer, Heidelberg (2010)
12. Roberts, M.G., Cootes, T.F., Adams, J.E.: Automatic Segmentation of Lumbar Vertebrae on Digitised Radiographs Using Linked Active Appearance Models. In: *Proc. Medical Image Understanding and Analysis*, pp. 120–124 (2006)
13. Roberts, M.G., Cootes, T.F., Adams, J.E.: Vertebral Morphometry: Semiautomatic Determination of Detailed Shape from Dual-energy X-ray Absorptiometry Images Using Active Appearance Models. *Investigative Radiology* 41, 849–859 (2006)
14. Roberts, M.G., Cootes, T.F., Pacheco, E., Oh, T., Adams, J.E.: Segmentation of Lumbar Vertebrae Using Part-Based Graphs and Active Appearance Models. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) *MICCAI 2009*. LNCS, vol. 5762, pp. 1017–1024. Springer, Heidelberg (2009)

Computer-Assisted Intramedullary Nailing Using Real-Time Bone Detection in 2D Ultrasound Images

Agnès Masson-Sibut^{1,2}, Amir Nakib¹, Eric Petit¹, and François Leitner²

¹ Laboratoire d'Images Signaux et Systèmes Intelligents (EA 3945), Université Paris Est Créteil, Créteil, France

² Research Center, Aesculap SAS, Echirolles, France

agnes.masson-sibut@bbraun.com, amir.nakib@u-pec.fr, eric.petit@u-pec.fr,
francois.leitner@bbraun.com

Abstract. In this paper, we propose a new method for bone surface detection in 2D ultrasound (US) images, and its application in a Computer Assisted Orthopaedic Surgery system to assist the surgeon during the locking of the intramedullary nail in tibia fractures reduction. It is a three main steps method: first, a vertical gradient is applied to extract potential segments of bone from 2D US images, and then, a new method based on shortest path is used to eliminate all segments that do not belong to the final contour. Finally, the contour is closed using least square polynomial approximation. The first validation of the method has been done using US images of anterior femoral condyles from 9 healthy volunteers. To calculate the accuracy of the method, we compared our results to a manual segmentation performed by an expert. The Misclassification Error (ME) is between 0.10% and 0.26% and the average computation time was 0.10 second per image.

Keywords: 2D ultrasound, bone surface, segmentation, Computer Assisted Surgery.

1 Introduction

In Computer Assisted Orthopaedic Surgery (CAOS) systems, the intra-operative image modality of choice is often Computed Tomography (CT) or fluoroscopy (X-rays projection). These image modalities are not completely safe for patients and users because they produce ionized radiations. Within the last decade, ultrasounds (US) became a valuable alternative for orthopaedic surgeons. Ultrasound devices are not too expensive, and portable. Also, ultrasound imaging can be used in real-time intra-operatively and it is non-invasive. However, the US images are difficult to analyze for the surgeon because of the high level of attenuation, shadow, speckle and signal dropouts [1].

In the literature, the extraction of the bone surface in US images was studied by Heger *et al.* [2]. The authors used an A-mode ultrasound pointer. The

probe was tracked mechanically to register the distal femur in total hip replacement. The A-mode of ultrasound probes consists in using the representation of the signal amplitude for one line in the tissue. Usually, the B-mode is preferred by surgeons, since the output is an image representing the brightness of the US response for several lines in the tissue. In CAOS systems, US imaging can be used to collect some sample landmarks on bone surface [3], or to perform intra-operative registration, extracting the full 3D model of an anatomical structure [4]. Manual segmentation of the bone surface in US images is highly operator dependent and time consuming [5]. Moreover, the thickness of the bone surface can reach 4 mm in some images [1], so manual segmentation can lead to an error higher than some millimeters. Foroughi *et al.* [6] developed an automatic segmentation method of bone surface in US images using dynamic programming. This method depends on a threshold value. The obtained average error was between 2.10 pixels to 2.67 pixels at the comparison between automatic and manual segmentation ; the average time of computation per image was 0.55 seconds.

In this paper, our main interest lies on the use of US images in computer assisted intramedullary nailing of tibia shaft fractures. If a surgeon chooses to reduce a tibia shaft fracture using an intramedullary nail, then he has to lock the nail in the bone. Normand *et al.* proposed to use some measures on the healthy symmetric tibia to assist the surgeon during the locking of the nail [7]. The authors noticed that the locking of the intramedullary nail without assistance can lead to important modifications in the orientation or the length of the broken tibia. They suggested using the healthy symmetrical tibia as model to assist the surgeon during the locking. Then, the broken bone could be reconstructed identical to its symmetrical. The proposed system used 3D positions of some anatomical landmarks (malleolus, trochlea, femoral condyles, ...) on both the broken and the healthy tibia to calculate their length and orientation, and the healthy tibia should not be injured. Then, the authors proposed to use the US probe as a percutaneous pointer (Fig. 1). Our work focuses on the development of a new method to extract the bone surface in real-time in order to find automatically anatomical landmarks.

The proposed method consists in three main steps. In the first step, a vertical gradient is applied to extract potential segments of the bone from 2D US images. In the second step, a new method based on a shortest path algorithm is used to eliminate all segments that do not belong to the final contour. Finally, the contour is closed using polynomial interpolation.

The rest of the paper is organized as follows: in Section 2, the proposed method is presented. Then, we show and discuss the obtained results in Section 3. The conclusion is in Section 4.

2 Proposed Method

Let $I : \Omega \subset \mathbb{N}^2 \rightarrow \mathcal{I} \subset \mathbb{N}$ be an image (two dimensional (2D) real function). Segmenting bone surface in I consists in extracting $\{P_i | i = 1, \dots, n\}$ a subset of contiguous points in I , where $P_i = (x_i, y_i) \in \Omega, \forall i = 1, \dots, n$. Considering

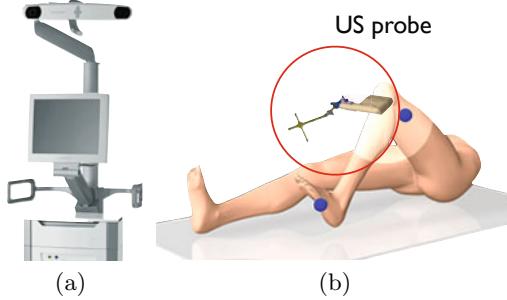


Fig. 1. Illustration of the CAOS system to assist the surgeon during the locking of the nail. (a) The station used to track the US probe. (b) Position of the patient during the US acquisition of anterior femoral condyles.

ultrasound properties of bones [1], we admit that $\forall(i, j) \in [1, n]^2$ such that $i \neq j$, we have $y_i \neq y_j$.

Then, the proposed segmentation method consists in three steps: in the first step, original images are filtered, a vertical gradient is computed, and an extraction of some potential segments of bone contour is performed. Then, the second step consists in characterizing these segments of contour, in order to eliminate those that *a priori* do not belong to the bone contour. Final step consists in closing the contour using least square polynomial approximation.

2.1 Preprocessing Step

Ultrasound images are highly textured, mainly with speckle. Then, the first sub-step of preprocessing is to apply a low-pass filter to the original image in order to eliminate noise and to strengthen interesting features. To find the best low pass filter, a frequency analysis of the different sequences by hand was done. We tested empirically several low-pass filters: Gaussian filters were too permissive but a good noise attenuation was found using a circular averaging filter. Once the filtered image I_s is computed, the vertical gradient is applied to it, and we denote the result image I_g (Fig. 2.(b)). The choice of the vertical gradient is motivated by ultrasound propagation properties, where the bone contours are mainly horizontal. It was shown in [1] that the bone contour is likely to lie in the top of the US bone response. Then, we only keep high values of the gradient.

To isolate the high values of the gradient, we need to threshold I_g (Fig. 2.(c)). In order to get a threshold that is independent of the image parameters (contrast, ...), we propose to use the cumulative histogram \mathcal{H}_{cum} of gradient values which is defined by:

$$\mathcal{H}_{cum}(x) = \sum_{i=1}^x h(x), \forall x = 1, \dots, N \Delta q \quad (1)$$

where $h(x)$ is the histogram value at the gradient level x of I_g , N is the number of levels in gradient values, and Δq is the step between two levels. We define the

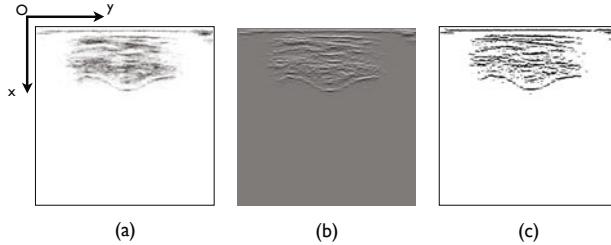


Fig. 2. Illustration of the preprocessing step method. In order to improve the visibility of US images, the grey levels are inverted. (a) Inverted initial image. (b) Gradient image. (c) Thresholded image.

optimal threshold (p) for I_g that expresses the percentage of low gradient values that have to stay at "0". Here, we define $p = 0.95$. Then,

$$I_{BW} = \begin{cases} 1 & \text{if } I_g > t \text{ where } t = \mathcal{H}_{cum}^{-1}(p \times \mathcal{H}_{cum}(N\Delta q)) \\ 0 & \text{otherwise} \end{cases} \quad (2)$$

Using properties of ultrasound imaging of bones [1], we can extract from I_{BW} a first subset of potential contour points $\{Q_i | i = 1, \dots, c\}$, where c is the number of columns in I . Considering that we can have at most one Q_i per column, contour points are denoted $Q_i = (x_i, i)$, $\forall i = 1, \dots, c$ in the rest of the paper. The subset of Q_i is built by taking the lowest non-zero point in each column of I_{BW} :

$$Q_i(I) = \begin{cases} argmax\{k | I_{BW}(k, i) \neq 0\} & \forall i = 1, \dots, c \\ 0, \text{ if } argmax = \emptyset \end{cases} \quad (3)$$

The next step consists in characterizing these points to determine whether or not they belong to the bone contour.

2.2 False Alarm Elimination

The subset of points $\{Q_i | i = 1, \dots, c\}$ (Fig. 3.(a)) are potentially part of the bone contour. To select those that belong to the bone contour, we consider them as segments by grouping contiguous points. Two points are considered to be contiguous if they belong to the same 3×3 neighborhood. In this step, we assume that segments smaller than 5 pixels correspond to noise, so they are eliminated. Those closer than 50 pixels to the top of the image are also eliminated because they are too close to the skin. For each segment k , where $k = 1, \dots, M$, the first point is designated by Q_{a_k} and the last point by Q_{b_k} , where a_k and b_k are the column of Q_{a_k} and Q_{b_k} respectively.

To define segments that belong to the bone contour, we define $\mathcal{G}(\mathcal{N}, \mathcal{E})$ as an oriented graph, using the M segments (Figure 4):

$$\begin{aligned} \mathcal{N} &= \{n_i | i = 1, \dots, 2M\} \\ &= \{a_k | k = 1, \dots, M\} \cup \{b_k | k = 1, \dots, M\} \end{aligned} \quad (4)$$

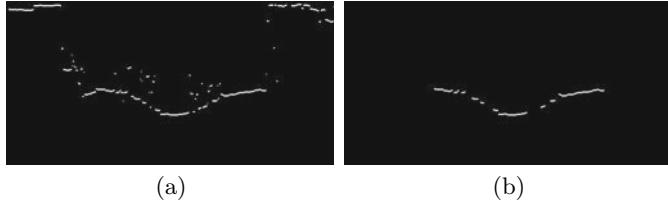


Fig. 3. Illustration of the false alarm elimination step. (a) Potential contour points before the false alarm elimination. (b) Remaining points after the Dijkstra algorithm.

is the set of all nodes in the graph, where the node index n_i is defined by:

$$\forall i \in [1, 2M], \quad n_i = \begin{cases} a_{\frac{i+1}{2}} & \text{if } i \text{ is odd} \\ b_{\frac{i}{2}} & \text{if } i \text{ is even} \end{cases} \quad (5)$$

We define also the set of edges in the graph as: $\forall (i, j) \in [1, 2M]^2$,

$$\mathcal{E}(i, j) = \begin{cases} \frac{1}{2}(b_{\frac{j}{2}} - a_{\frac{i}{2}}) & \text{if } i \text{ is odd and } j = i + 1 \\ \|Q_{b_{\frac{j}{2}}} - Q_{a_{\frac{j+1}{2}}}\| & \text{if } i \text{ is even, } j \text{ is odd, and } i < j < \min(2M, i + 6) \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

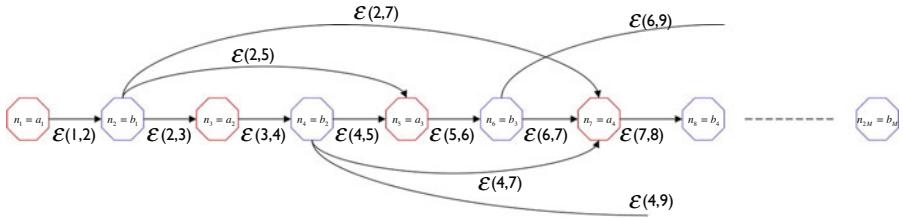


Fig. 4. The construction of the graph \mathcal{G} . We distinguish nodes called "start of segments" which are the a_k nodes and the nodes called "end of segments" which are the b_k nodes.

The graph \mathcal{G} contains two types of nodes: nodes *start-of-segments*, which are $\{a_k | k = 1, \dots, M\}$ with a single child (b_k), and the weight of corresponding edges is $\frac{1}{2}(b_{\frac{k}{2}} - a_{\frac{k}{2}})$, and, the nodes *end-of-segments* which are $\{b_k | k = 1, \dots, M\}$ with at most three children which are $\{a_l | l = k+1, \dots, \min(k+3, M)\}$, and the weight of corresponding edges are $\|Q_{b_{\frac{k}{2}}} - Q_{a_{\frac{l+1}{2}}}\|$. The intra-segment edges values are penalized to enforce inter-segment distance as driving force and are considered in the computation of the global shortest path.

Then, Dijkstra's algorithm [8] is used to find the shortest path between the node n_1 and the node n_{2M} in the oriented graph \mathcal{G} . An illustration of the obtained result after the application of Dijkstra algorithm is presented Figure 3.(b).

2.3 Contour Closure

The closure of the contour is performed by a polynomial approximation using least square method. The approximation is calculated on the m points that belong to the final set of segments. The degree of the polynomial has been defined empirically by $R = 10 < m$ in order to have a unique solution. An illustration of the closure is given in Figure 5.

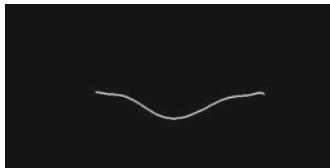


Fig. 5. Illustration of the contour closure

3 Results

To illustrate the effectiveness of the method, different tests were performed on several series of US images. The probe and the beamformer were provided by Telemed (Vilnius, Lithuania), and the acquisition software was developed by Aesculap SAS (Echirolles, France) using the Software Development Toolkit provided by Telemed. The acquisition parameters were a frequency at 7 MHz and two focus points at 20 mm and 34 mm. The acquisition protocol consists in locking the probe under the patella, and performing a scan of the femoral condylar region rotating the probe up and down. To test the proposed method, a database for 29 men and women, from 24 to 61 year-old, was used. The results presented here are on 230 US images from 9 representative volunteers from this database.

To qualify the final segmentation, the Root Mean Square Error (RMSE) and Misclassification Error (ME) [9] measures were used to compare the obtained results to an expert's segmentation. After outliers elimination, the maximal value of the RMSE is around 7 pixels ($\approx 0.8\text{mm}$) which is quite good considering that the acceptable error is around 2mm in a CAOS applications. The ME values between 0.10% and 0.26% illustrate the accuracy of the method comparing to the expert's segmentation. The algorithm was tested on Matlab R2008a and runs at 10-15 images per seconds, which can be considered as a real-time in our context.

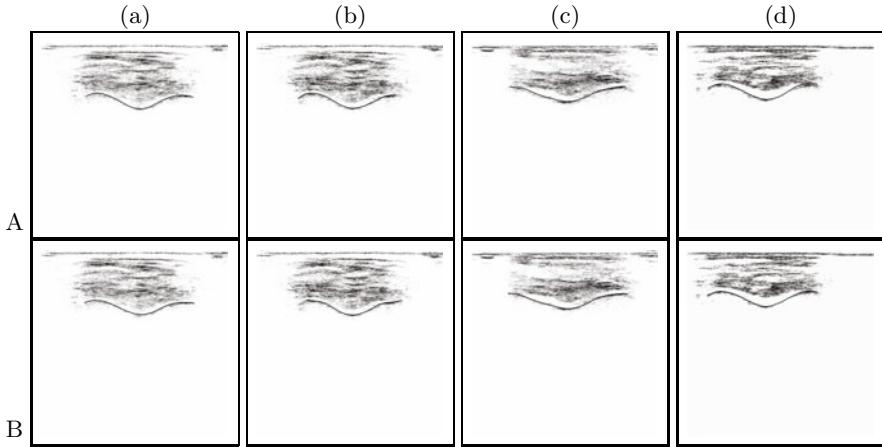


Fig. 6. Comparison between the expert's segmentation and the automatic segmentation for images from 4 different volunteers. A.(a)(b)(c)(d) Results of the expert's segmentation. B.(a)(b)(c)(d) Results of the automatic segmentation.

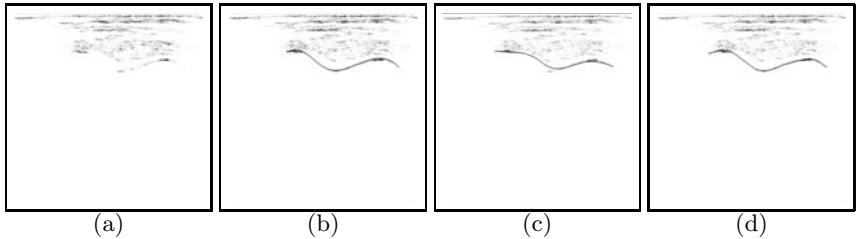


Fig. 7. Illustration of the comparison between segmentation result using an active contour, and the proposed method. (a) Original US image. (b) US image segmented by an expert. (c) US image segmented using an active contour model. (d) US image segmented using the proposed method.

Figure 6 shows some results: the first line represents the results of the expert's segmentation, and the second line represents the results given by the automatic segmentation. The difference between them is not visually significant.

The obtained results of the proposed method were compared to those obtained by the snake based method described in [10]. The comparison is illustrated in Figure 7. The RMSE value between the segmentation result of the snake based method and the manual segmentation is 14.4 pixels (11.5mm).

4 Conclusion

In this paper, a new method of condyles detection in ultrasound images was presented. The method took into account both ultrasounds, and bones properties

to have the best final result. We tested the method on 230 images, with the same acquisition parameters. First results are very promising, especially in our context of application: its use in a CAOS system to assist the surgeon during intramedullary nailing of tibia shaft fractures.

Nevertheless, automatic interpretation of US images is non trivial: images of the same organ can vary a lot depending on the patient, and on the set of parameters used for the acquisition. In future work, a study on the settings of the acquisition parameters will be done.

In work under progress, to improve the quality of the segmentation, a geometric a priori of the condyles contour is added.

References

1. Jain, A.K., Taylor, R.H.: Understanding bone responses in B-mode ultrasound images and automatic bone surface extraction using a Bayesian probabilistic framework. In: Proceedings of International Conference SPIE Medical Imaging, vol. 5373, pp. 131–142. Spie, Bellingham (2004)
2. Heger, S., Portheine, F., Ohnsorge, J., Schkommodau, E., Radermacher, K.: User-interactive registration of bone with A-mode ultrasound. IEEE Engineering in Medicine and Biology Magazine 24, 85–95 (2005)
3. Beek, M., Abolmaesumi, P., Luenam, S., Sellens, R.W., Pichora, D.R.: Ultrasound-guided percutaneous scaphoid pinning: Operator variability and comparison with traditional fluoroscopic procedure. In: Larsen, R., Nielsen, M., Sporrings, J. (eds.) MICCAI 2006. LNCS, vol. 4191, pp. 536–543. Springer, Heidelberg (2006)
4. Zhang, Y., Rohling, R., Pai, D.: Direct surface extraction from 3D freehand ultrasound images. In: Proceedings of the Conference on Visualization 2002, p. 52. IEEE Computer Society, Boston (2002)
5. Barratt, D.C., Penney, G.P., Chan, C.S.K., Slomczykowski, M., Carter, T.J., Edwards, P.J., Hawkes, D.J.: Self-calibrating 3D-ultrasound-based bone registration for minimally invasive orthopedic surgery.. IEEE Transactions on Medical Imaging 25, 312–323 (2006)
6. Foroughi, P., Boctor, E., Swartz, M.J., Taylor, R.H., Fichtinger, G.: Ultrasound Bone Segmentation Using Dynamic Programming. In: 2007 IEEE Ultrasonics Symposium Proceedings, pp. 2523–2526. IEEE, New York (2007)
7. Normand, J., Harisboure, A., Leitner, F., Pinzuti, J., Dehoux, E., Masson-Sibut, A.: Experimental navigation for bone reconstruction. In: 10th Annual Meeting of The International Society for Computer Assisted Orthopaedic Surgery Proceedings, Versailles, France (2010); Poster 39
8. Dijkstra, E.W.: A note on two problems in connexion with graphs. Numerische Mathematik 1, 269–271 (1959)
9. Sezgin, M., Sankur, B.: Survey over image thresholding techniques and quantitative performance evaluation. Journal of Electronic Imaging 13, 146–165 (2004)
10. Kass, M., Witkin, A., Terzopoulos, D.: Snakes: Active contour models. International Journal of Computer Vision 1, 321–331 (1988)

Multi-Kernel Classification for Integration of Clinical and Imaging Data: Application to Prediction of Cognitive Decline in Older Adults

Roman Filipovych¹, Susan M. Resnick², and Christos Davatzikos¹

¹ Department of Radiology, University of Pennsylvania, Philadelphia, PA, USA

² Laboratory of Personality and Cognition, Biomedical Research Center/04B317,
251 Bayview Blvd., Baltimore, MD 21224

Abstract. Diagnosis of neurologic and neuropsychiatric disorders typically involves considerable assessment including clinical observation, neuroimaging, and biological and neuropsychological measurements. While it is reasonable to expect that the integration of neuroimaging data and complementary non-imaging measures is likely to improve early diagnosis on individual basis, due to technical challenges associated with the task of combining different data types, medical image pattern recognition analysis has been largely focusing solely on neuroimaging evaluations. In this paper, we explore the potential of integrating neuroimaging and clinical information within a pattern classification framework, and propose that the multi-kernel learning (MKL) paradigm may be suitable for building a multimodal classifier of a disorder, as well as for automatic identification of the relevance of each information type. We apply our approach to the problem of detecting cognitive decline in healthy older adults from single-visit evaluations, and show that the performance of a classifier can be improved when neuroimaging and clinical evaluations are used *simultaneously* within a MKL-based classification framework.

Keywords: Multi-Kernel Learning (MKL), Normal aging, MRI.

1 Introduction

Image-based high-dimensional pattern classification has gained considerable attention, and has begun to provide tests of high sensitivity and specificity on an individual patient basis, in addition to characterizing group differences [7,6,11]. One of the major advantages of pattern recognition methods is their ability to capture multivariate relationships among various anatomical regions for more effective characterization of group differences.

Despite the advances of image-based pattern recognition, clinical observations still form the basis for the diagnosis in neurologic and neuropsychiatric disorders. Moreover, besides imaging evaluations, there are a number of non-imaging measures that are of significance in a variety of studies. For example, the potential alternative markers associated with aging can be biological [2,12], genetic [10], or cognitive [5]. While the potential of a computational approach that would

integrate disparate types of information (*e.g.*, structural, functional, clinical, genetic, *etc.*) is obvious, technical challenges associated with such an approach often prevented joint use of the available alternative measures.

Several attempts have been made recently in the direction of integrating different types of imaging or non-imaging information for pattern classification in the studies of aging. Zhang *et al.* [20] integrate MRI, PET modalities with CSF markers via a weighted combination of multiple kernels, which provided improvements in the problem of discriminating Alzheimer's disease (AD) (or Mild Cognitive Impairment (MCI)) and healthy controls. In a similar problem of classifying AD and healthy subjects, Hinrichs *et al.* [9] employed a multi-kernel learning (MKL) approach to integrate different imaging and non-imaging data. In [4], a computational imaging marker of AD-like atrophy was combined with CSF to predict MCI to AD conversion. At the same time, studies that would investigate possible benefits of integrating imaging and clinical evaluations for prediction of cognitive decline in healthy subjects are absent.

The prospect of finding potential treatments of AD makes it critical to identify biomarkers for very early stages of cognitive decline. As the diagnosis of MCI and AD involves comprehensive clinical observations, it is tempting to use cognitive evaluations when predicting aging-related cognitive decline. There are, however, several challenges associated with the reliability of cognitive measures. While cognitive measures may be significantly different in cognitively declined populations as compared to healthy controls, these measures may not be able to predict cognitive decline at baseline when the decline is not yet evident. As the result of the low predictive power of cognitive measures at baseline, as well as due to the associated significant noise, a number of followup evaluations are typically needed to detect cognitive decline. Considering that brain structure may actually precede reduction in cognitive function by at least several years [13], it is reasonable to expect that single-visit imaging evaluations together with the respective cognitive measures can jointly provide richer information for the detection of cognitive decline.

In this paper, we describe a general MKL-based framework that integrates imaging and clinical information for classification. Our approach consists of an image processing pipeline and a MKL component that combines disparate data for classification. The application focus of this paper is in predicting cognitive decline in healthy older adults by combining MRI and a cognitive assessment test, where our method allows inference about longitudinal outcomes based on the analysis of single-visit imaging and cognitive evaluations.

2 Background

Support Vector Machines (SVM) [19] have been shown to provide high classification accuracy, and are among the most widely used classification algorithms in the brain MRI classification studies [7,11]. SVM project the data into a high-dimensional space, and find the classification function as the separating hyperplane with the largest margin, where the margin is the distance from the separating hyperplane to the closest training examples.

Multi-Kernel Learning (MKL) [18] extends the theory of SVM by allowing different kernel functions to represent subsets of features (e.g., MRI, cognitive evaluations). Given a set of points $\mathcal{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$, and their respective labels $\{y_1, \dots, y_n\}$, the MKL problem for K kernels can be formulated as follows:

$$\begin{aligned} & \min_{\beta_k, \mathbf{w}_k, b, \xi} \frac{1}{2} \left(\sum_{k=1}^K \beta_k \|\mathbf{w}_k\|_2 \right)^2 + C \sum_{i=1}^n \xi_i \\ \text{s.t. } & y_i \left(\sum_{k=1}^K \beta_k \mathbf{w}^T \varphi_k(\mathbf{x}_i) + b \right) \geq 1 - \xi_i, \quad \xi_i \geq 0, \quad \forall i = 1, \dots, n; \end{aligned} \quad (1)$$

where the slack variables ξ_i are introduced to allow some amount of misclassification in the case of non-separable classes, constant C implicitly controls the tolerable misclassification error, the kernel functions $\varphi_k(\mathbf{x})$ map the original data into a high-dimensional, possibly infinitely-dimensional, space, and $\beta = (\beta_1, \dots, \beta_K)$ are the subkernel weights. The sparsity of the kernel combinations is controlled by a constraint on the subkernel weights, where the commonly used sparse constraint is $\|\beta\|_1 = 1$, and the typical non-sparse constraint is $\|\beta\|_2 = 1$. The task of the MKL optimization problem is then to find subkernel weights while simultaneously maximizing the margin for the training data.

3 Classification of Imaging and Clinical Evaluations

Figure 1 presents the diagram of our approach, and uses the task of integrating structural MRI and cognitive evaluations as an example. The main components of our approach are: (1) tissue density estimation; (2) ROI features extraction; and (3) integration of image measurements and clinical evaluations via MKL.

Tissue density estimation. All MR images are preprocessed following a mass-preserving shape transformation framework [3]. Gray matter (GM) and white matter (WM) tissues are segmented out from each skull-stripped MR brain image by a brain tissue segmentation method [14]. Each tissue-segmented brain image is then spatially normalized into a template space, by using a high-dimensional image warping method [16]. The total tissue mass is preserved in each region

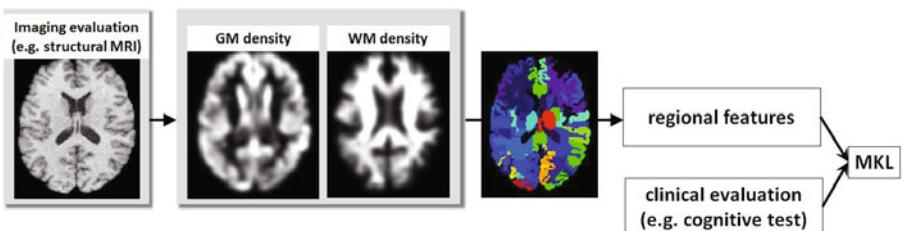


Fig. 1. Diagram of our approach (example of integrating structural MRI and cognitive evaluations). The main steps include tissue density estimation, ROI features extraction, and integration of imaging and cognitive evaluations via MKL.

during the image warping, and tissue density maps are generated in the template space. These tissue density maps give a quantitative representation of the spatial distribution of tissue in a brain, with brightness being proportional to the amount of local tissue volume before warping.

Extracting regional features. The original brain images, and the respective tissue density maps, have very high dimensionality which makes it difficult to discover patterns in the resulting high-dimensional space. By registering brain images to a common template with a predefined number of labeled anatomical regions of interest (ROIs), and by calculating mean tissue density at each ROI, the dimensionality of the original data can be reduced to a manageable size. We use a template image with 101 manually labeled ROIs, and compute average GM and WM tissue densities at ROIs as the image features.

MKL classification. For the purpose of integrating imaging and clinical evaluations we use one kernel for imaging features, and another kernel for the respective clinical evaluations. We use the publicly available implementation of ℓ_2 -norm MKL [17], and consider Gaussian kernels with σ_i and σ_c representing kernel widths for imaging features and clinical measures, respectively.

4 Results

4.1 Materials and Evaluation

Dataset. We analyzed a population of 127 healthy individuals from the Baltimore Longitudinal Study of Aging (BLSA) [15] which has been following a set of older adults with annual and semi-annual imaging and clinical evaluations. In this paper we focus on MRI evaluations of the BLSA as the imaging component of our analysis. In conjunction with each imaging evaluation, every individual's cognitive performance was evaluated on tests of mental status and memory. We selected the following three measures for our analysis: the immediate free recall score (sum of five immediate recall trials) on the California Verbal Learning Test (CVLT) [5], the total number of errors from the Benton Visual Retention Test (BVRT) [1], and the total score from the Mini-Mental State Exam (MMSE) [8].

Defining cognitively stable and declining groups. While the individual cognitive evaluations are often noisy and unreliable, one can identify trends of cognitive decline by considering rates of change in cognitive evaluations over time. We formed the cognitively stable labeled subset from 25 subjects who had the highest slopes of CVLT, and 25 subjects with the lowest CVLT slope values were assigned into the labeled cognitively declining subset. The slope of the CVLT score represents the rate of cognitive decline, and lower slopes of the score indicate higher rates of decline. The remaining 77 subjects were unlabeled.

Evaluation. In order to assess the classification performance of our approach, we adopted a leave-one-out (LOO) evaluation scheme (Figure 2). At each run of the LOO evaluation we removed one subject from the labeled set as the test

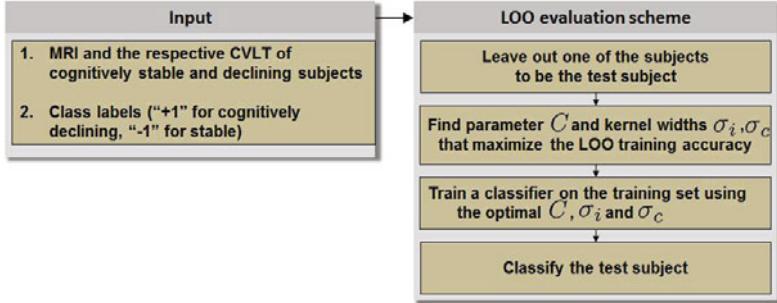


Fig. 2. LOO evaluation scheme (example of integrating structural MRI and CVLT)

subject. The remaining subjects formed the training set and the classifier was trained on the training data. The free parameters of the classifier (*e.g.*, C , kernel widths) were identified as the ones that yielded the highest LOO classification accuracy on the *training* set. After the classifier was trained on the training set, it was applied to the test subject to obtain the subject's test label.

4.2 Classification of Single-Visit Evaluations

In our first experiment, we classified baseline (*i.e.*, first-visit) evaluations following the leave-one-out scheme in Figure 2. The distributions of age in cognitively stable and cognitively declining subpopulations had means 65.8 ± 6.3 years and 70.4 ± 7.0 years, respectively, with the age of cognitively declining individuals being slightly higher than in the stable subjects ($p = 0.02$). The MKL classifier with a single kernel is equivalent to the SVM classifier, and yielded classification LOO accuracy of 58.0% when using only the MRI information, and 66.0% when using only CVLT at baseline. At the same time, by integrating MRI and CVLT at baseline, we were able to achieve classification accuracy of 74.0%. Table 1 summarizes the classification performance at baseline using imaging evaluations, cognitive evaluations, and the combination of both. The subkernel weights β estimated by MKL using all labeled subjects were 0.884 and 0.467 for the MRI and CVLT kernels, respectively.

Similarly, we assessed the ability of our approach to detect cognitive decline using last-visit evaluations. The mean ages in cognitively stable and cognitively declined populations during the last visits were 73.4 ± 7.6 and 78.2 ± 6.4 , respectively, with age in cognitively declining individuals being slightly higher than in

Table 1. Classification of first-visit evaluations

Kernels	Accuracy	Sensitivity	Specificity
MRI	58.0%	52.0%	64.0%
CVLT	66.0%	68.0%	64.0%
MRI+CVLT	74.0%	76.0%	72.0%

Table 2. Classification of last-visit evaluations

Kernels	Accuracy	Sensitivity	Specificity
MRI	70.0%	76.0%	64.0%
CVLT	88.0%	88.0%	88.0%
MRI+CVLT	88.0%	88.0%	88.0%

the stable subjects ($p = 0.02$). Table 2 summarizes the classification performance for last-visit evaluations. The CVLT-only classifier at last visits performed on a par with the MKL classifier that integrated MRI and CVLT. The subkernel weights estimated by MKL using the last-visit evaluations of all labeled subjects were 0.162 and 0.987 for the MRI and CVLT kernels, respectively.

As expected, the performance of the MKL classifier in the task of discriminating cognitively stable and cognitively declining individuals was significantly better for relatively older individuals (*i.e.*, 88.0% during last visits vs. 74.0% during first visits).

4.3 Biomarker of Cognitive Decline

Next, we investigated whether the classifier trained on the last visit scans can predict cognitive decline during earlier evaluations. For a given test subject with a set of longitudinal evaluations $\mathbf{x}_1, \dots, \mathbf{x}_t$, and the MKL classifier estimated from the last visit evaluations of the training subjects, we obtained the values of the classification function $\mathcal{F}(\mathbf{x}_1), \dots, \mathcal{F}(\mathbf{x}_t)$, where $\mathcal{F}(\mathbf{x}) = \sum_{k=1}^K \beta_k \mathbf{w}_k^T \varphi_k(\mathbf{x}) + b$. The value of the classification function for the subject's evaluation \mathbf{x} reflects the presence of brain phenotypic as well as cognitive pattern inherent to cognitive decline, with larger values of the classification function indicating higher similarity with “imaging-cognitive” pattern of decline. The plot in Figure 3 shows sensitivity and specificity of the classifier tested at every given year of evaluation, where the year of evaluation for an individual is defined relative to the year of the individual's first visit. Note, that not all subjects may have undergone evaluation at any given year. Indeed, apart from the first and the third year of evaluation, none of the evaluation years witnessed evaluations performed for all 50 labeled subjects. Moreover, some subjects had as few as four evaluations, while some had as many as eleven. Consequently, classification results shown in the plot in Figure 3 were obtained for different, although overlapping, sets of subjects. For example, only 34 subjects had evaluation during their 8-th year. As a result, the classification performance of the classifier trained on the last-

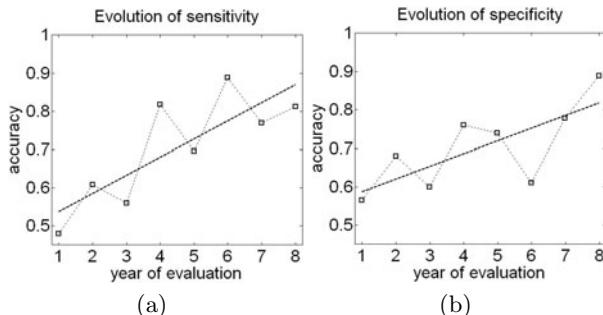


Fig. 3. Sensitivity (a) and specificity (b) of the classifier trained on last-visit evaluations and applied to earlier visits

visit evaluations of the training subjects and applied to all evaluations of the test subject may not be directly comparable for any two years of evaluation. Nonetheless, the trends in Figure 3 suggest that the classification performance of the classifier noticeably improves with subjects' age. The performance of the MKL classifier trained on the *last-visit* evaluations and applied to the very early evaluations is somewhat surprising. In particular, the sensitivity and specificity of the classifier trained on the last-visit evaluations is very low when applied to the first-visit evaluations (*i.e.*, sensitivity and specificity at year one in Figure 3(a,b)). At the same time, the results in Section 4.2 show that the classifier that is specifically trained on the *first-visit* evaluations can predict cognitive decline based on first-visit evaluations with significantly higher accuracy. This may suggest that different classifiers may be needed for prediction of cognitive decline from the evaluations of different age groups.

4.4 Analysis of Individuals with Uncertain Trends of Decline

As we described in the Materials, the trends of cognitive decline in 77 out of 127 individuals were not clear, and the subjects could not have been reliably assigned into one of the two labeled groups. We analyzed the ability of the MKL classifier trained on the last-visit evaluations of the labeled subjects to detect cognitive decline in the subjects with weak trends of decline. After the MKL classifier was trained on the last-visit evaluations to discriminate between cognitively stable and cognitively declining labeled sets, we obtained values of the classification function for every evaluation of the 77 unlabeled individuals.

Figure 4(a) shows correlation between the classification values and the rate of change in CVLT for different years of evaluation. In general, correlation between the classification values and rate of change in cognitive performance increases with age, which is expected given that the MKL classifier was trained on the last-visit evaluations of the labeled individuals. Additionally, Figures 4(b) and 4(c) show evolution of correlation between the classification values and BVRT and

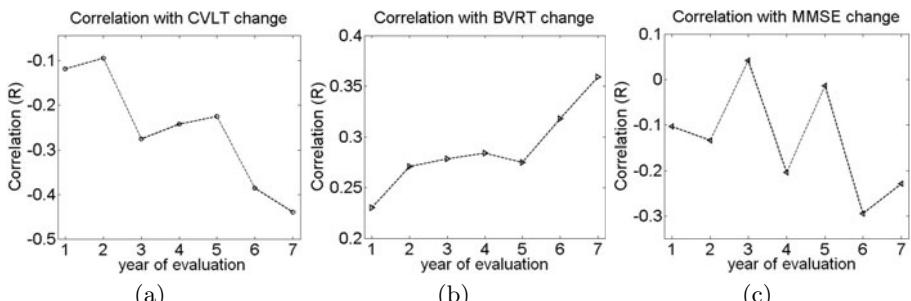


Fig. 4. Correlation between the value of the classification function and the *rate of change* in cognitive evaluations for specific evaluation years. (Lower values of CVLT and MMSE, as well as higher values of BVRT, indicate worse cognitive performance).

MMSE, respectively. While the increase in correlation with BVRT is evident, increase in correlation with MMSE during later evaluations is less obvious, which reflects the fact that MMSE is typically more noisy than CVLT and BVRT.

5 Conclusion

In this paper, we presented a pattern classification framework for integration of imaging and non-imaging evaluations. Our method involves an image preprocessing and feature extraction protocol, and employs MKL methodology to integrate imaging and non-imaging features. The application focus of our approach was in prediction of cognitive decline in healthy older adults, where we used MKL to integrate single-visit structural neuroimaging and cognitive evaluations. Our results suggest that, while neither MRI nor CVLT individually carry sufficient information to predict cognitive decline based on a single evaluation, they allow us to achieve promising prediction accuracy when considered jointly. Our proposed approach is general and can potentially be used for integration of other types of neuroimaging and non-imaging data. In particular, we are planning to further explore the problem of predicting cognitive decline in older adults by integrating structural MRI, PET, and other cognitive, as well as genetic, evaluations.

Acknowledgments. This research was supported in part by the Intramural Research Program of the NIH, National Institute on Aging (NIA), and R01-AG14971, N01-AG-3-2124, N01-AG-3-2124.

References

1. Benton, A.: Revised Visual Retention Test. The Psych. Corp., New York (1974)
2. Bouwman, F.H., van der Flier, W.M., Schoonenboom, N.S.M., van Elk, E.J., Kok, A., Rijmen, F., Blankenstein, M.A., Scheltens, P.: Longitudinal changes of CSF biomarkers in memory clinic patients. *Neurology* 69(10), 1006–1011 (2007)
3. Davatzikos, C., Genc, A., Xu, D., Resnick, S.M.: Voxel-based morphometry using the ravens maps: methods and validation using simulated longitudinal atrophy. *Neuroimage* 14(6), 1361–1369 (2001)
4. Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q.: Prediction of mci to ad conversion, via mri, csf biomarkers, and pattern classification. *Neurobiology of Aging* (2010) (in press, corrected proof)
5. Delis, D., Kramer, J., Kaplan, E., Ober, B.: California Verbal Learning Test - Research Edition. The Psychological Corporation, New York (1987)
6. Duchesne, S., Bocti, C., De Sousa, K., Frisoni, G.B., Chertkow, H., Collins, D.L.: Amnestic mci future clinical status prediction using baseline mri features. *Neurobiol Aging* 31(9), 1606–1617 (2010)
7. Fan, Y., Batmanghelich, N., Clark, C.M., Davatzikos, C.: Spatial patterns of brain atrophy in mci patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *NeuroImage* 39(4), 1731–1743 (2008)
8. Folstein, M.F., Folstein, S.E., McHugh, P.R.: "mini-mental state". a practical method for grading the cognitive state of patients for the clinician. *J. Psychiatr. Res.* 12(3), 189–198 (1975)

9. Hinrichs, C., Singh, V., Xu, G., Johnson, S.C.: Predictive markers for ad in a multi-modality framework: An analysis of mci progression in the adni population. *NeuroImage* 55(2), 574–589 (2011)
10. Ji, Y., Permanne, B., Sigurdsson, E.M., Holtzman, D.M., Wisniewski, T.: Amyloid beta40/42 clearance across the blood-brain barrier following intra-ventricular injections in wild-type, apoe knock-out and human apoe3 or e4 expressing transgenic mice. *J. Alzheimers Dis.* 3(1), 23–30 (2001)
11. Kloppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack, C.R., Ashburner, J., Frackowiak, R.S.J.: Automatic classification of mr scans in alzheimer's disease. *Brain* 131(3), 681–689 (2008)
12. de Leon, M., Mosconi, L., Li, J., De Santi, S., Yao, Y., Tsui, W., Pirraglia, E., Rich, K., Javier, E., Brys, M., Glodzik, L., Switalski, R., Saint Louis, L., Pratico, D.: Longitudinal csf isoprostane and mri atrophy in the progression to ad. *Journal of Neurology* 254, 1666–1675 (2007)
13. Petersen, R., Jack Jr., C.: Imaging and biomarkers in early alzheimer's disease and mild cognitive impairment. *Clin. Pharmacol. Ther.* 84(4), 438–441 (2009)
14. Pham, D.L., Prince, J.L.: Adaptive fuzzy segmentation of magnetic resonance images. *IEEE Trans. Med. Imaging* 18(9), 737–752 (1999)
15. Resnick, S.M., Pham, D.L., Kraut, M.A., Zonderman, A.B., Davatzikos, C.: Longitudinal magnetic resonance imaging studies of older adults: A shrinking brain. *J. Neurosci.* 23(8), 3295–3301 (2003)
16. Shen, D., Davatzikos, C.: Hammer: Hierarchical attribute matching mechanism for elastic registration. *IEEE Trans. Med. Imag.* 21(11), 1421–1439 (2002)
17. Sonnenburg, S., Rätsch, G., Henschel, S., Widmer, C., Behr, J., Zien, A., de Bona, F., Binder, A., Gehl, C., Franc, V.: The shogun machine learning toolbox. *J. Mach. Learn. Res.* 99, 1799–1802 (2010)
18. Sonnenburg, S., Rätsch, G., Schäfer, C., Schölkopf, B.: Large scale multiple kernel learning. *J. Mach. Learn. Res.* 7, 1531–1565 (2006)
19. Vapnik, V.N.: The nature of statistical learning theory. Springer-Verlag New York, Inc., New York (1995)
20. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D.: Multimodal classification of alzheimer's disease and mild cognitive impairment. *NeuroImage* 55(3), 856–867 (2011)

Automated Selection of Standardized Planes from Ultrasound Volume

Bahbibi Rahmatullah^{1,*}, Aris Papageorghiou², and J. Alison Noble¹

¹ Institute of Biomedical Engineering, Dept. of Engineering Science, University of Oxford
{bahbibi.rahmatullah, alison.noble}@eng.ox.ac.uk}

² Nuffield Dept of Obstetrics and Gynaecology, John Radcliffe Hospital, University of Oxford
aris.papageorghiou@obs-gyn.ox.ac.uk

Abstract. The search for the standardized planes in a 3D ultrasound volume is a hard and time consuming process even for expert physicians. A scheme for finding the standardized planes would be beneficial in advancing the use of volumetric ultrasound for clinical diagnosis. In this paper, we propose a new method to automatically select the standard plane from the fetal ultrasound volume for the application of fetal biometry measurement. To our knowledge, this is the first study in the fetal ultrasound domain. The method is based on the AdaBoost learning algorithm and has been evaluated on a set of 30 volumes. The experimental results are promising with a recall rate of 91.29%. We believe this will increase the accuracy and efficiency in patient monitoring and care management in obstetrics, specifically in detecting growth restricted fetuses.

Keywords: AdaBoost, Ultrasound, Standardized plane, Detection.

1 Introduction

Comprehensive ultrasound examination during pregnancy includes standard fetal biometric measurements, which are primarily used to estimate the gestational age of the fetus, to track fetal growth pattern, to estimate fetal weight and to detect abnormalities. Fetal growth assessment is used to identify IUGR (intra-uterine growth restriction) which is prevalent in millions of babies who are born every year with low birth weight because of IUGR and/or prematurity, resulting in significant short- and long term morbidity and mortality [1]. Growth restricted fetuses have poorer neonatal outcomes and it is now well recognized that developmental delay associated with IUGR leads to significant health care and developmental problems during childhood and most likely in adult life [2]. Recognition of the serious risks associated with IUGR has elevated its diagnostic importance.

Fetal biometry is determined from standardized ultrasound planes taken from the fetal head, abdomen and thigh. Thus, the acquisition of standard image planes where these measurements are taken from is crucial to allow for accurate and reproducible biometric measurements, and also to minimize inter- and intra-observer variability. Currently, the fetal biometry is measured from images produced by 2D ultrasound scans. One major criticism of the 2D ultrasound scan is the operator bias and the reproducibility issues [3]. There is always the risk of incorrect measurement taken by

* Corresponding Author.

the operator because of image acquisition in the less appropriate scan plane [4]. Alternatively, a new scan after a significant time lapse is non-ideal due to the missed opportunity to take the measurement at the specific fetal age to assess growth and also the cost factor to the physician and patient. 3D ultrasound offers to reduce this uncertainty since a clinician can repeat measurement for that particular fetal age scan anytime by finding the appropriate plane from the volumetric data. However, the search for standardized planes from the 3D volume is a painstaking task and a time consuming process even for expert physicians.

Previous research conducted on finding standardized planes in ultrasound images has mostly confined to the domain of 3D echocardiography. Slice to volume registration methods were used for finding standard anatomical views for rest and stress images [5]. The registration technique benefits from the standard protocol in cardiac ultrasound for the placement of ultrasound probe on patient. However, it is not suitable for fetal ultrasound because of the variability introduced by the inconsistent positioning of the probe on the patient body and the fetal position during the scan.

In the field of fetal ultrasound, development of automated image analysis method is so far limited to the problem of segmenting specific anatomical structures for the biometric measurement estimation. The work and the literature review in [6] presents an excellent summary and advancement of automated methods in fetal ultrasound. To the best of our knowledge, no one has looked at the optimality of the plane used for taking the biometric measurement from and our work is the first one to look at automating the plane standardization in fetal biometry ultrasound. We choose to focus on the standardization of fetal abdominal plane since the abdominal circumference (AC) measurement taken from this plane is currently the most important measure in assessing fetal growth and IUGR. The variability in AC measurement is broader compared to the other two fetal biometry measures [7]. However the methodology presented in this paper is general and applicable for standardization of other image planes in ultrasound and other imaging modality.

In fetal abdominal ultrasound image, the standard plane is determined based on the presence of two important anatomical landmarks which are the stomach bubble (SB) and the umbilical vein (UV). The plane containing these two landmarks is described in [8] and used in constructing the widely used chart for abdominal circumference size by [9]. Our main work in this paper is to automate the detection of these two landmarks (SB and UV) and use this information to select a range of standardized image planes from the 3D volume. Fig. 1 shows the different planes acquired at different locations in a fetal abdominal ultrasound volume.

2 Plane Selection Method

Our proposed method involves user labeled training data, discriminant classifiers of local regions, and weighted combination of those local classifier results. The overall approach can be divided into three stages: 1) feature extraction, 2) training of weak classifiers and automatic feature selection, and 3) selection of the most probable slices using the classifier scores plot.

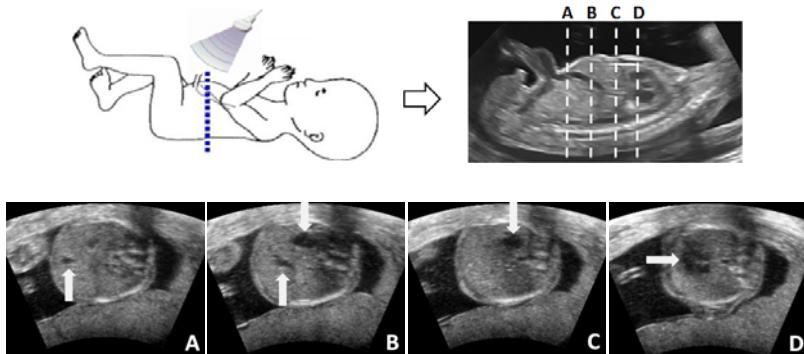


Fig. 1. Illustration of different slices acquired from a 3D volume at different positions on the fetus. Slice A is too low indicated by umbilical vein (UV) (arrow) that is near the abdomen wall. Slice B is an optimal slice with stomach bubble (SB) (top arrow) and UV (lower arrow) at correct position. Slice C has only SB (arrow) and Slice D is too high near the heart (arrow).

2.1 Feature Extraction

To construct the model for selecting the image slices, we extracted the image regions that consist of SB and/or UV from the standardized slices selected by our experts. We experimented with two separate detection model:

- a. *Method 1*: One Combined Trained Classifier (1CTC) for detecting both SB and UV in one region.
- b. *Method 2*: Two Separately Trained Classifiers (2STC) each for detecting SB and UV and the scores from each classifier are combined as the final hypothesis.

The appearance of the image region is represented with features derived from the Haar wavelets [10]. The Haar features had been proven to effectively encode the domain knowledge for different types of visual patterns including objects in ultrasound images [11]. Extraction is achieved with high speed of computation due to the use of the integral image [12]. No separate feature selection process is needed since the most representative features are automatically selected during the training process.

2.2 Learning Algorithm

We choose to explore AdaBoost [13], an intuitive and proven effective method in object detection problem, such as face detection in natural images [12] and in medical images, such as detection of landmark points in echocardiogram [11] and lung bronchovascular anatomy in CT data [14]. Advantages of AdaBoost algorithm are that it has no parameters to tune other than the number of iterations and the feature selection process is done simultaneously during the training process, requiring no additional experiments.

AdaBoost originated from the idea of ‘Boosting’, a process of forming strong hypothesis through linear combination of weak ones. In the context of object detection which is a binary classification task, a weak hypothesis can be represented as the weak classifiers that are derived from a pool of extracted features. These

classifiers are called ‘weak’ because even the best classification function is not expected to classify the data well, they only need to classify correctly the examples in more than 50% of the cases. AdaBoost uses modification of the weight distributions based on the previous classification error in order to focus on the more difficult samples, thus driving down the classification error. The AdaBoost implementation for training our classifier model is outlined in Fig. 2. The first five features selected by the algorithm for each detection model are displayed in Fig. 3.

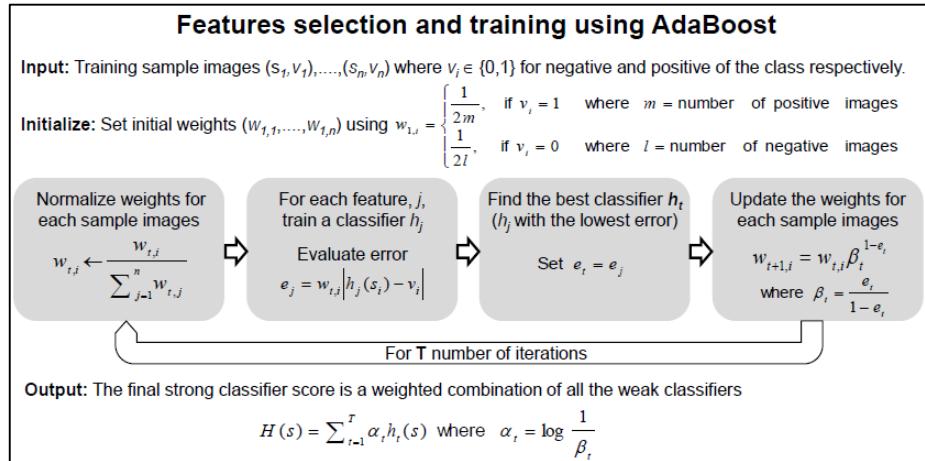


Fig. 2. Framework for feature selection and classifier training using AdaBoost

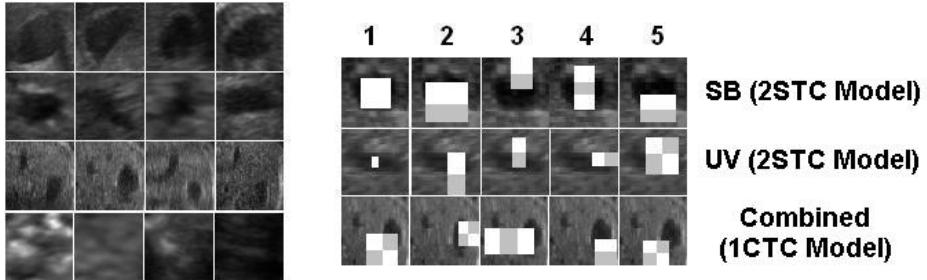


Fig. 3. Examples of 4 image classes used for training (left), i.e., stomach bubble (SB) (first row), umbilical vein (UV) (second row), SB and UV combined in one window (third row) and background (bottom row). The first five selected Haar features by AdaBoost for each trained models are shown imposed on some example images from the training set (right). The Haar features are calculated by subtracting the total intensity in white regions with total intensity in grey regions.

2.3 Slice Selection

Our input volume is regarded as slices of coronal (y-axis) planes. By using the sliding window method, features indicated by the weak classifiers are extracted from each

image patch that covers the entire image plane. The maximum score from all the patches in a particular slice is used to denote the probability of SB and/or UV detection for that image slice. For Method 1 (1CTC), the final score is calculated by dividing the product of both scores by their sum. The plots showing the scores from both methods for all slices in a test volume sample are shown in Fig. 4. The standardized slices are selected from the peak with the global maximum point using empirically determined threshold values for all testing volumes (0.77 and 0.42 for Method 1 (1CTC) and Method 2 (2STC), respectively). The flowchart in Fig. 5 gives the summary of our automated slice selection method implementation.

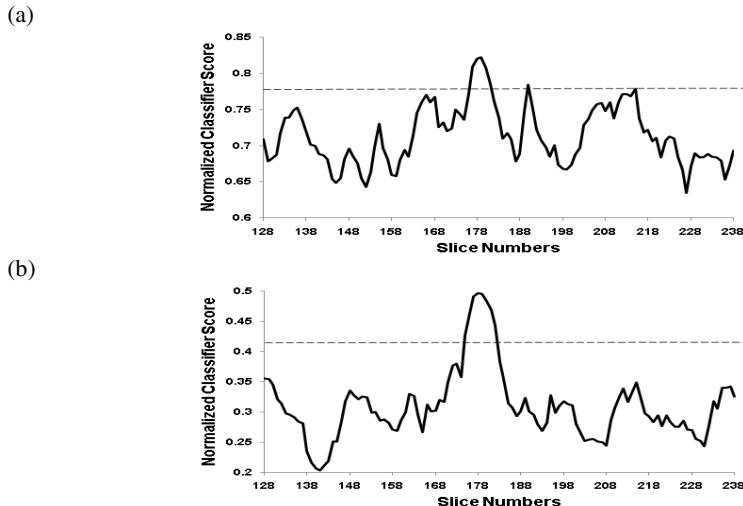


Fig. 4. Graphs showing the normalized classifier scores for each slices in a sample volume for (a) Method 1 (1CTC) and (b) Method 2 (2STC). The dotted lines represent the threshold levels employed in our method.

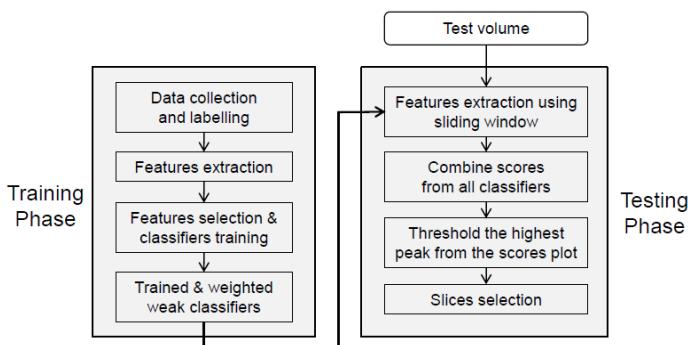


Fig. 5. Flowchart showing the implementation of the training and testing phase

3 Datasets and Experiments

3.1 Data Acquisition

Fetal abdominal volumes for this work were randomly selected from various ultrasound scans performed on different women. Scans were performed using a Philips HD9 ultrasound machine with a 3-7MHz 3D probe and saved in MVL file format. The MVL file were converted to a RAW file with metadata header using RegisterAndConvert[©] software with output resolution of 0.33mm in x, y, and z directions. Average size of volume was 312x268x244. Volumes were divided into three sets without overlap: 10 volumes for training, five volumes for algorithm validation and 30 volumes for testing. Fetuses were scanned at 20-28 weeks of gestational ages because most of fetal abdominal measurements are performed within this timeframe. The image slices from each volumes were annotated as standard optimal planes after consultation with trained sonographer.

3.2 Results and Performance Analysis

The experiment was conducted using MATLAB 7.6 on a Pentium Xeon® 3.4 GHz machine. In our validation experiment, we used different number of weak classifiers (50, 100, 150, and 200) to classify 100 regions containing SB (or UV) and 100 background region samples. The ROC analysis presented in Fig. 6 indicates that 100 weak classifiers are sufficient for our model since there is no significant increase in classification accuracy by adding more weak classifiers after that. This indicates that the first 100 weak classifiers capture most of the distinguishing information for our model.

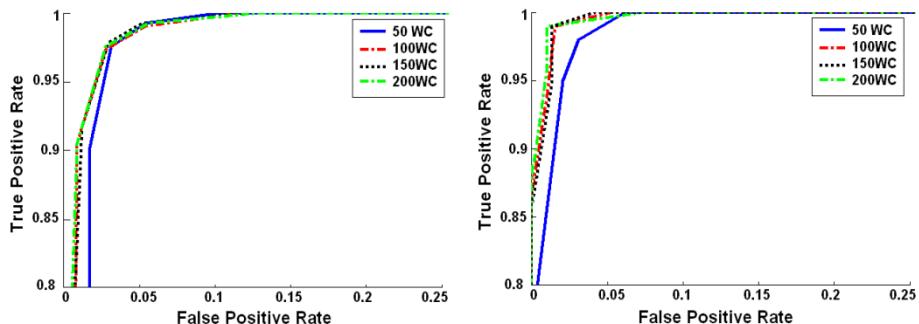


Fig. 6. ROC analysis for effect of increasing the number of weak classifiers (WC) (50, 100, 150, and 200) in SB detection (*left*) and UV detection (*right*)

Table 1. The performance of two different methods used in standardized slice selection from fetal abdominal volume

Selection method	Average Precision (%)	Average Recall (%)
Method 1 (1CTC)	51.93	64.76
Method 2 (2STC)	76.29	91.29

Two evaluation criterias were used to evaluated the performance of the automated method:

$$\text{Precision, } P = \frac{\text{True Positive}}{\text{True Positive} + \text{False Positive}} \quad (1)$$

$$\text{Recall, } R = \frac{\text{True Positive}}{\text{True Positive} + \text{False Negative}} \quad (2)$$

Table 1 shows a comparison between the results of the two methods that we used in training the model to detect SB and UV in image slice and their execution time. 100 weak classifiers were used for each model based on the result observed in Fig. 6

Clearly the strategy to have separate classifier for each object (SB and UV) in Method 2 (1STC) outperforms the combined object detection approach in Method 1 (1CTC) although Method 2 is slower due to having two separate classifiers that operate on the image slices serially. Most of the slices were correctly selected by our proposed method as indicated by the high recall percentage.

All of the False Positive images selected by Method 1 contained both the SB and UV in correct positions but may have not been selected as optimal by the expert because of a stricter criteria imposed. We plan to investigate the inter-user agreement by including more experts to have a more accurate evaluation of the methods.

4 Conclusion

We have developed a novel selection method for finding standardized image planes from fetal abdominal ultrasound volumes. To our best knowledge, this is the first study to look at automating the image plane selection from fetal ultrasound volume driven by the detection of anatomical features found in the image using a machine learning method. The approach to use separate classifiers that are trained using AdaBoost learning algorithm for detecting SB and UV enables the method to correctly select most of the manually labeled slices by experts. The method has an average recall rate of 91.29% and precision of 76.29% as compared to simultaneous detection of both object using the same classifier that resulted in a lower recall and precision rate of 51.93% and 64.76%, respectively.

Our results validate the concept of automated fetal sonography and potentially can become a tool for improving the efficiency of 3D ultrasound imaging making it more appealing to the clinicians and obstetricians. Time needed to complete an ultrasound examination could be reduced and an off-line examination can be repeated easily with higher accuracy in measurement, thereby resulting in better clinical management of high risk patients especially those identified with IUGR.

Acknowledgments. The authors gratefully acknowledge the financial support of the Ministry of Higher Education (MOHE), Malaysia and Sultan Idris Education University (UPSI), Malaysia. We also would like to thank Dr Ippokratis Sarris for the helpful discussion.

References

1. Lawn, J.E., Cousens, S., Zupan, J.: Million neonatal deaths: When? Where? Why? *Lancet* 365, 891–900 (2005)
2. Barker, D.J.P.: Adult consequences of fetal growth restriction. *Clinical Obstetrics and Gynecology* 49, 270–283 (2006)
3. Chan, L.W., Fung, T.Y., Leung, T.Y., Sahota, D.S., Lau, T.K.: Volumetric (3D) imaging reduces inter- and intraobserver variation of fetal biometry measurements. *Ultrasound in Obstetrics and Gynecology* 33, 447–452 (2009)
4. Elliott, S.T.: Volume ultrasound: The next big thing? *British Journal of Radiology* 81, 8–9 (2008)
5. Leung, K.Y.E., et al.: Sparse registration for three-dimensional stress echocardiography. *IEEE Transactions on Medical Imaging* 27, 1568–1579 (2008)
6. Carneiro, G., Georgescu, B., Good, S., Comaniciu, D.: Detection and measurement of fetal anatomies from ultrasound images using a constrained probabilistic boosting tree. *IEEE Transactions on Medical Imaging* 27, 1342–1355 (2008)
7. Hadlock, F.P., Deter, R.L., Harrist, R.B., Park, S.K.: Fetal abdominal circumference as a predictor of menstrual age. *American Journal of Roentgenology* 139, 367–370 (1982)
8. Campbell, S., Wilkin, D.: Ultrasonic measurement of fetal abdomen circumference in the estimation of fetal weight. *British Journal of Obstetrics and Gynaecology* 82, 689–697 (1975)
9. Chitty, L.S., Altman, D.G., Henderson, A., Campbell, S.: Charts of fetal size: 3. Abdominal measurements. *British Journal of Obstetrics and Gynaecology* 101, 125–131 (1994)
10. Oren, M., Papageorgiou, C., Sinha, P., Osuna, E., Poggio, T.: Pedestrian detection using wavelet templates, pp. 193–199 (1997)
11. Karavides, T., Leung, K.Y.E., Paclik, P., Hendriks, E.A., Bosch, J.G.: Database guided detection of anatomical landmark points in 3D images of the heart, pp. 1089–1092. Rotterdam (2010)
12. Viola, P., Jones, M.J.: Robust Real-Time Face Detection. *International Journal of Computer Vision* 57, 137–154 (2004)
13. Freund, Y., Schapire, R.E.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* 55, 119–139 (1997)
14. Ochs, R.A., et al.: Automated classification of lung bronchovascular anatomy in CT using AdaBoost. *Medical Image Analysis* 11, 315–324 (2007)

Maximum Likelihood and James-Stein Edge Estimators for Left Ventricle Tracking in 3D Echocardiography

Engin Dikici¹ and Fredrik Orderud²

¹ Norwegian University of Science and Technology, Trondheim, Norway

² GE Vingmed Ultrasound, Oslo, Norway

Abstract. Accurate and consistent detection of endocardial borders in 3D echocardiography is a challenging task. Part of the reason for this is that the trabeculated structure of the endocardial boundary leads to alternating edge characteristics that varies over a cardiac cycle. The maximum gradient (MG), step criterion (STEP) and max flow/min cut (MFMC) edge detectors have been previously applied for the detection of endocardial edges. In this paper, we combine the responses of these edge detectors based on their confidences using maximum likelihood (MLE) and James-Stein (JS) estimators. We also present a method for utilizing the confidence-based estimates as measurements in a Kalman filter based left ventricle (LV) tracking framework. The effectiveness of the introduced methods are validated via comparative analyses among the MG, STEP, MFMC, MLE and JS.

1 Introduction

3D echocardiography has enabled real-time, non-invasive and low cost acquisition of volumetric images of the LV. The problem of automatic detection and tracking of heart chambers in ultrasound images has received considerable attentions lately [1,2]. However, the accurate detection of the endocardial borders remains a challenging task. This is partially due to the trabeculated structure of the endocardial borders, which leads to alternating edge characteristics over a cardiac cycle.

One approach for the LV detection is to use a Kalman filter based tracking framework to update a deformable model based on the edge measurements. In an early work by Blake et al., Kalman filtering was used for tracking B-spline models deformed in an affine shape space [3]. In their study, object boundaries were determined by selecting the gradient maxima (MG) of image intensity profiles. Later, this framework was utilized with a principal component analysis based shape space for the LV tracking in 2D ultrasound by Jacob et al.[4]. This study employed a local-phase edge detector [5] for the edge measurements, and reported visually enhanced results compared to the maximum gradient method. Orderud et al. utilized an extended Kalman filter to track deformable subdivision surfaces in 3D image data sets [2]. The latter work used a step criterion (STEP) [6] for the detection of endocardial edges. More recently, Dikici et al. applied the max flow

/ min cut algorithm (MFMC) for the detection of endocardial edges in Kalman tracking framework [7]. Their study provided comparative analyses representing the shortcomings of STEP and MFMC methods at end-diastole (ED) and end-systole (ES), and accordingly proposed a hybrid edge detector.

Ensemble methods combine the responses of multiple predictors for answering a novel instance. They have been shown to be effective as the resulting classifiers are often more accurate than the individual classifiers making up the ensemble [8]. The concept can be applied on a variety of tasks including the edge detection in image processing. Konishi et al. used the statistical inference to combine responses of multiple edge detectors (e.g. intensity gradient, the Laplacian of a Gaussian, filterbanks of oriented filter pairs) to produce highly accurate edge detectors [9]. The fusion entropy was utilized to integrate decisions from different detectors in order to improve the edge detection accuracy in [10].

This paper proposes two confidence-based endocardial edge detection methods employing (1) maximum likelihood (MLE) and (2) James-Stein (JS) estimators, where the MG, STEP and MFMC are the base predictors in the ensemble. The edge detection responses of the base predictors are combined using their estimation variances, which are learned from a training dataset. The effectiveness of the introduced methods are validated via comparative analyses among MG, STEP, MFMC, MLE and JS.

2 Tracking Framework

The tracking framework is built around a deformable subdivision model parametrized by a set of control vertices and their associated displacement direction vectors. Model deformations are handled by a composite transform, where local shape deformations are obtained by moving control vertices in the subdivision model together with a global transformation that translates, rotates and scales the whole model.

A manually constructed Doo-Sabin surface is used to represent the endocardial borders. This model consists of 20 control vertices that are allowed to move in the surface normal direction to alter the shape. The edge detection is conducted from a set of approximately 500 surface points, spread evenly across the endocardial surface.

The tracking framework consists of five separate stages, namely the (1) state prediction, (2) evaluation of tracking model, (3) edge measurements, (4) measurement assimilation, and (5) measurement update. In this study, the stages are identical as in [2], and therefore not covered. The edge detection methods used during the *edge measurements* stage are further investigated.

3 Edge Detection Methods

The edge detection process is performed by first extracting N 1D intensity profiles (I_1, I_2, \dots, I_N) centered around the surface points p_i and oriented in the surface normal directions n_i . The total number of samples in each profile, K ,

and the distance between consecutive samples are determined empirically. $I_{i,k}$ is used for referring to the intensity value of the i^{th} intensity profile's k^{th} sample. Edge detection methods, processing intensity profiles to estimate endocardial border positions, are described in the following subsections.

3.1 Maximum Gradient Edge Detector (MG)

The intensity profile I_i is convolved with a Gaussian kernel G to create a smoother intensity profile. Then, a gradient profile for the smoothed profile is computed by using the forward-difference approximation. The position of the maximum of the gradient profile is selected as the edge index. The measurement noise is set inversely proportional with the maximum gradient. For each profile, the edge index is determined as:

$$s_i = \operatorname{argmax}_k (| [I * G]_{i,k} - [I * G]_{i,k+1} |). \quad (1)$$

3.2 Step Criterion Edge Detector (STEP)

STEP assumes that the intensity profile I_i forms a transition from one intensity plateau to another. It calculates the heights of the two plateaus for each index value, and selects the index with the lowest sum of squared differences between the criteria and the image data. For each profile, the edge index will then be determined as:

$$s_i = \operatorname{argmin}_k \sum_{t=0}^{k-1} \left(\left(\frac{1}{k} \sum_{j=0}^{k-1} I_{i,j} \right) - I_{i,t} \right)^2 + \sum_{t=k}^{K-1} \left(\left(\frac{1}{K-k} \sum_{j=k}^{K-1} I_{i,j} \right) - I_{i,t} \right)^2 \quad (2)$$

The measurement noise is set inversely proportional with the height difference between the plateaus.

3.3 Max Flow/Min Cut Edge Detector (MFMC)

The max flow/min cut algorithm can be used for finding the global optima of a set of energy functions including,

$$E(f) = \sum_{v \in V} D_v(f_v) + \sum_{(v,y) \in \text{Edges}} Q_{v,y}(f_v, f_y). \quad (3)$$

The optimization process seeks a labeling function f that assigns binary values to the nodes that are defined under a set V , distinguishing inside of the LV cavity ($f = 1$) from the outside ($f = 0$). The classification is constrained by *data penalty* D_v , and *interaction potential* $Q_{v,y}$ functions. D_v penalizes the labeling of v based on the predefined likelihood function, whereas $Q_{v,y}$ penalizes the labeling discontinues between the neighboring nodes v and y . The problem of finding the optimal edges is formulated as *eqn-3* in MFMC edge detector. Initially, a graph with nodes representing the profile samples is created. The source and sink terminal nodes are appended to the node-set, and connected with the

nodes corresponding to the first and the last samples of the intensity profiles respectively. These connections are called the *t-links* and have infinite weights. The nodes corresponding to (1) the consecutive samples of the same profile, and (2) the same index samples of the neighboring profiles, are connected by undirected edges called the *n-links*. The weight of an n-link connecting the nodes v and y is calculated as:

$$\text{weight}(v, y) = C \times \exp\left(\frac{-(I_v - I_y)^2}{2\sigma^2}\right), \quad (4)$$

where I_v and I_y refers to the intensity values at the associated profile samples and C is a constant. After the graph is created, the maximum flow / minimum cut between the source and sink nodes are found using the push-relabel algorithm. The resulting cut defines the edge positions for all intensity profiles simultaneously. The reverse of the flow amount is utilized as the measurement noise in the Kalman filter [7].

3.4 Maximum Likelihood (MLE) and James-Stein (JS) Estimators

The estimation accuracies of MG, STEP and MFMC methods change depending on the cardiac cycle position. Informally, MFMC detects the endocardial edges better at ED, while MG and STEP are more precise at ES [7]. The estimates of these methods can be combined using a statistical learning approach for generating better estimates; the confidences of the detectors for a given instance determine their temporal weights.

For a given edge detector, estimated edge indices for the model can be represented as:

$$S_\zeta = [s_{1,\zeta}, s_{2,\zeta}, \dots, s_{N,\zeta}], \quad (5)$$

where $s_{i,\zeta}$ is the estimated edge index for the i^{th} intensity profile, and $\zeta \in [0 : ES, 1 : ED]$ gives the cardiac cycle position. Using a similar notation, the correct edge indices for the model can be defined as $\lambda_\zeta = [\nu_{1,\zeta}, \nu_{2,\zeta}, \dots, \nu_{N,\zeta}]$. The mean of the signed error for S_ζ can be calculated simply as,

$$Err_\zeta = \frac{1}{N} \sum_{i=1}^N (s_{i,\zeta} - \nu_{i,\zeta}). \quad (6)$$

The bias and variance properties of the estimator can be estimated using a training dataset of size B using,

$$Bias_\zeta = E[Err_\zeta] \approx \frac{1}{T} \sum_{b=1}^B Err_\zeta^{(b)}, \quad (7)$$

$$\sigma_\zeta^2 = E[(Err_\zeta - E[Err_\zeta])^2] \approx \frac{1}{T} \sum_{t=1}^T (Err_\zeta^{(b)} - Avg_b(Err_\zeta))^2, \quad (8)$$

where $Err_{\zeta}^{(b)}$ gives the signed error for the cardiac cycle position ζ for the b^{th} training sample. Hence, an unbiased edge detector μ_{ζ} for the i^{th} intensity profile can be defined as $\mu_{\zeta} = s_{i,\zeta} - Bias_{\zeta}$. Assuming that $\mu_{\zeta} = \nu_{i,\zeta} + \varepsilon_{\zeta}$, where $\varepsilon_{\zeta} \sim N(0, \sigma_{\zeta}^2)$, pair $\langle \mu_{\zeta}, \sigma_{\zeta}^2 \rangle$ represents the variance coupled estimate for a given instance. This type of predictor is called the *probability-variance predictor*. As our ensemble includes 3 base estimators (MG, STEP and MFMC), their estimates for the i^{th} intensity profile are given by $\langle \mu_{\zeta,(1)}, \sigma_{\zeta,(1)}^2 \rangle$, $\langle \mu_{\zeta,(2)}, \sigma_{\zeta,(2)}^2 \rangle$ and $\langle \mu_{\zeta,(3)}, \sigma_{\zeta,(3)}^2 \rangle$. Then, the maximum likelihood estimator (MLE) for the i^{th} intensity profile at time ζ can be found by,

$$\hat{\nu}_{i,\zeta,(MLE)} = \frac{\sum_{j=1}^3 \frac{\mu_{\zeta,(j)}}{\sigma_{\zeta,(j)}^2}}{\sum_{j=1}^3 \frac{1}{\sigma_{\zeta,(j)}^2}}. \quad (9)$$

$\hat{\nu}_{i,\zeta,(MLE)}$ is an unbiased estimator as,

$$E[\hat{\nu}_{i,\zeta,(MLE)}] = E\left[\frac{\sum_{j=1}^3 \frac{\mu_{\zeta,(j)}}{\sigma_{\zeta,(j)}^2}}{\sum_{j=1}^3 \frac{1}{\sigma_{\zeta,(j)}^2}}\right] = \frac{\sum_{j=1}^3 \frac{E[\mu_{\zeta,(j)}]}{\sigma_{\zeta,(j)}^2}}{\sum_{j=1}^3 \frac{1}{\sigma_{\zeta,(j)}^2}} = \nu_{i,\zeta}. \quad (10)$$

The variance of the $\hat{\nu}_{i,\zeta,(MLE)}$ is given by,

$$\sigma^2[\hat{\nu}_{i,\zeta,(MLE)}] = \frac{1}{\sum_{j=1}^3 \frac{1}{\sigma_{\zeta,(j)}^2}}. \quad (11)$$

MLE simply weights the estimators with the inverse of their estimation variances, which can be thought of as their confidences. Furthermore, the detector variances are found dependent on time, making them *classifier+instance variances* [11]. Due to this property, the estimator can assign different weights to an edge detector for different cardiac cycle positions.

The James Stein estimator (JS) is a biased estimator with a non-compact point distribution function. However, JS dominates MLE considering the expected prediction error with a quadratic loss function [12]. For applying JS, $\nu_{i,\zeta}$ is assumed to be distributed as $N(0, \tau_{\zeta}^2)$. JS seeks for the mean of the posterior distribution $E[\lambda_{\zeta} | \mu_{\zeta}, \sigma_{\zeta}^2]$, which depends on the unknown parameter τ_{ζ}^2 . Therefore, the method first estimates τ_{ζ}^2 using empirical Bayes estimation [12]. The steps of this well published estimation are omitted in this paper. The resulting JS estimator for $\nu_{i,\zeta}$ is given by,

$$\hat{\nu}_{i,\zeta,(JS)} = \left(1 - \frac{\sum_{j=1}^3 \frac{1}{\sigma_{\zeta,(j)}^2}}{\left(\sum_{j=1}^3 \frac{\mu_{\zeta,(j)}}{\sigma_{\zeta,(j)}^2}\right)^2}\right)^+ \hat{\lambda}_{i,\zeta,(MLE)}, \quad (12)$$

where $(\alpha)^+ = \max(\alpha, 0)$ [11].

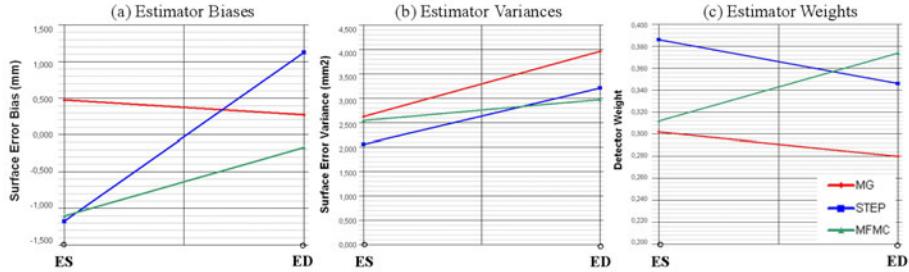


Fig. 1. (a) Surface error biases, (b) surface error variances, and (c) the corresponding weights of the edge detectors over the cardiac cycle.

4 Results

A set of 18 apical 3D echocardiography recordings were acquired using a Vivid 7 ultrasound scanner (GE Vingmed Ultrasound, Norway) and a matrix array transducer. The endocardial border segmentations for the recordings were performed by a medical expert using a semi-automatic segmentation tool (4D AutoLVQ, GE Vingmed Ultrasound, Norway). Next, the recordings were randomly divided into training and testing datasets including 12 and 6 recordings respectively. MG, STEP and MFMC edge detectors were each employed in connection to the existing Kalman tracking framework for the training dataset. For each method, the bias and variance properties were estimated for the ED and ES frames. Then, these properties are interpolated to all cardiac cycle using a linear interpolation. Figure 1 shows the estimated surface error bias, variance, and the corresponding weights for the edge detectors.

For the testing, MG, STEP and MFMC estimator biases were eliminated using the time varying biases estimated from the training dataset. MLE and JS methods were set to combine the non-biased MG, STEP and MFMC methods with the varying weights represented in Figure 1-c. (1) The mean square error, averaging the signed surface error squares, (2) the mean absolute error, averaging the unsigned surface errors, and (3) the mean of the volume percentage errors were computed for MG, STEP, MFMC, MLE and JS methods using the testing dataset (see Figure 2).

During the training and testing, a handcrafted Doo-Sabin endocardial model consisting of 20 control points was used as the LV model. Edge measurements were performed in 528 intensity profiles evenly distributed across the endocardial model. Each profile consisted of 30 samples, spaced 1 mm apart. For MG and STEP, normal displacement measurements that were significantly different from their neighbors were discarded as outliers. The tracking framework is implemented in C++, and processed each frame in 6.8ms with MG, 7.5ms with STEP, 78ms with MFMC, 80ms with MLE, and 81ms with JS when executed on a 2.80 GHz Intel Core 2 Duo CPU.

5 Discussion and Conclusion

We have introduced MLE and JS methods that utilize statistical inference for the endocardial edge detection. To our knowledge, the idea of using MLE and JS to combine multiple edge detectors in medical imaging context is novel. Comparative analyses showed that the proposed methods lead to improved surface and volumetric measurements (see Figure 3 for the color coded surface error maps of a sample case). Furthermore, the estimator weights computed analytically using the statistical learning are consistent with the weights derived empirically in [7].

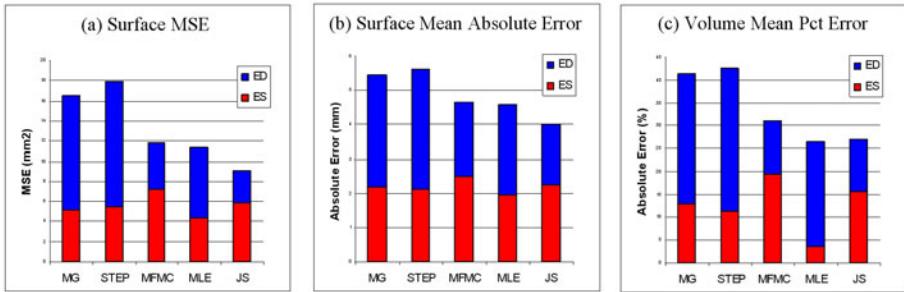


Fig. 2. (a) Surface mean square error, (b) surface mean absolute error, and (c) volume mean percentage error values for the Kalman tracking framework using MG, STEP, MFMC, MLE and JS edge detectors

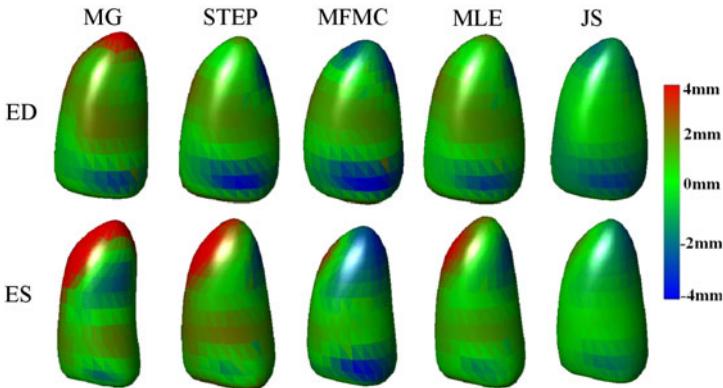


Fig. 3. The signed surface errors for ED (the upper row) and ES (the lower row) phases are shown for a case. 4mm over-estimation is red, 4mm under-estimation is blue, 0mm no-error is light green.

The proposed approach can be further improved in the following areas:

1. The bias and variances properties of the estimators are calculated only at ED and ES frames, then linearly interpolated to all cardiac cycle in this

- study. The properties could be calculated on ≥ 3 points of the cardiac cycle for generating higher order estimator weight curves.
2. The estimator properties are marginalized over the endocardial surface in this study. More explicitly, the bias and variance properties are assumed to be constant over the endocardial wall (i.e. the estimator bias at ED does not vary from apex to midwall). This restriction could be eliminated by parameterizing the statistical properties over the cardiac surface, or simply at a selected set of landmarks.
 3. Three base predictors, which are assumed to be uncorrelated, are employed in this study. As proved in [13], an ideal ensemble consists of highly accurate classifiers that disagree as much as possible. Hence, the ensemble size could be increased while factoring in the detector correlations in a future work.

References

1. Yang, L., Georgescu, B., Zheng, Y., Meer, P., Comaniciu, D.: 3d ultrasound tracking of the left ventricles using one-step forward prediction and data fusion of collaborative trackers. In: Proc. IEEE Conf. Computer Vision and Pattern Recognition (2008)
2. Orderud, F., Rabben, S.I.: Real-time 3d segmentation of the left ventricle using deformable subdivision surfaces. In: CVPR (2008)
3. Blake, A., Isard, M.: Active Contours: The Application of Techniques from Graphics, Vision, Control Theory and Statistics to Visual Tracking of Shapes in Motion. Springer-Verlag New York, Inc., Secaucus (1998)
4. Jacob, G., Noble, J.A., Mulet-Parada, M., Blake, A.: Evaluating a robust contour tracker on echocardiographic sequences. Medical Image Analysis 3, 63–75 (1999)
5. Venkatesh, S., Owens, R.: On the classification of image features. Pattern Recogn. Lett. 11, 339–349 (1990)
6. Rabben, S.I., Torp, A.H., Støylen, A., Slørdahl, S., Bjørnstad, K., Haugen, B.O., Angelsen, B.: Semiautomatic contour detection in ultrasound m-mode images. Ultrasound in Medicine & Biology 26, 287–296 (2000)
7. Dikici, E., Orderud, F.: In: Graph-Cut Based Edge Detection for Kalman Filter Based Left Ventricle Tracking in 3D+ T Echocardiography, pp. 1–4 (2010)
8. Opitz, D., Maclin, R.: Popular ensemble methods: An empirical study. Journal of Artificial Intelligence Research 11, 169–198 (1999)
9. Konishi, S., Yuille, A.L., Coughlan, J.M., Zhu, S.C.: Statistical edge detection: Learning and evaluating edge cues. IEEE Transactions on Pattern Analysis and Machine Intelligence 25, 57–74 (2003)
10. Li, J., Jing, X.: Edge detection based on decision-level information fusion and its application in hybrid image filtering. In: 2004 International Conference on Image Processing, ICIP 2004, vol. 1, pp. 251–254 (2004)
11. Lee, C.H., Greiner, R., Wang, S.: Using query-specific variance estimates to combine bayesian classifiers. In: ICML 2006, pp. 529–536. ACM, New York (2006)
12. Efron, B., Morris, C.: Stein’s estimation rule and its competitors—an empirical bayes approach. Journal of the American Statistical Association 68, 117–130 (1973)
13. Krogh, A., Vedelsby, J.: Neural network ensembles, cross validation, and active learning. In: Advances in Neural Information Processing Systems, pp. 231–238. MIT Press, Cambridge (1995)

A Locally Deformable Statistical Shape Model

Carsten Last¹, Simon Winkelbach¹, Friedrich M. Wahl¹,
Klaus W.G. Eichhorn², and Friedrich Bootz²

¹ Institut fuer Robotik und Prozessinformatik, TU Braunschweig, Germany

² Klinik und Poliklinik fuer HNO-Heilkunde/Chirurgie, UKB Bonn, Germany

c.last@tu-braunschweig.de

Abstract. Statistical shape models are one of the most powerful methods in medical image segmentation problems. However, if the task is to segment complex structures, they are often too constrained to capture the full amount of anatomical variation. This is due to the fact that the number of training samples is limited in general, because generating hand-segmented reference data is a tedious and time-consuming task. To circumvent this problem, we present a Locally Deformable Statistical Shape Model that is able to segment complex structures with only a few training samples at hand. This is achieved by allowing a unique solution in each contour point. Unlike previous approaches, trying to tackle this problem by partitioning the statistical model, we do not need predefined segments. Smoothness constraints ensure that the local solution is restricted to the space of feasible shapes. Very promising results are obtained when we compare our new approach to a global fitting approach.

1 Introduction and Related Work

Since their introduction by Cootes and Taylor [6] in 1995, *Statistical Shape Models* (SSMs) play an important role in medical image segmentation problems. However, one major problem in using statistical models for medical image segmentation tasks is that they often do not capture the full amount of anatomical variation, especially when the task is to segment complex structures, like e.g. the paranasal sinuses. This is because the limited amount of training data available is often not sufficient to estimate the high-dimensional probability distribution of feasible shapes. Thus, the resulting SSM is often over-constrained and not able to capture fine anatomical details [9].

Much effort has therefore been spent on making statistical models more flexible, giving them the ability to account for variations not present in the training data. Cootes and Taylor [4] tried to enlarge the training set by using vibrational modes, obtained from finite element methods, to generate a batch of synthetic samples from each real training sample. Later, they simplified their approach by directly modifying some elements in the covariance matrix of their trained SSM [5], hence obtaining more flexibility of neighboring contour elements. A comparison of various other approaches can be found in [12]. Without doubt, all these approaches enhance the flexibility of the SSM, but the main disadvantage is that they introduce synthetic variations *not* explained by the training shapes.

Another possibility to circumvent the drawback of over-constrained SSMs is to loosen the model constraints during the segmentation process. For instance, Shang and Dössel [16] use the SSM in a coarse-to-fine segmentation framework. In the last iteration of their approach, the use of the SSM is shut down in order to being able to fully adapt to the local image structures. Other approaches (e.g. [5], [19], and [17]) perform a trade-off between fitting prominent local image structures and sticking with the shapes explained by the SSM. Evidently, the shortcoming with theses approaches is that the final segmentation result is not required to be in conformance with the SSM.

The most promising approach to deal with the limited amount of training data is to partition the shapes and calculate their statistics individually in each segment (e.g. [3], [20]). Davatzikos *et al.* [7] and later on Nain *et al.* [13] further extended this by using the wavelet transform for a hierarchical partition of the training shapes. The problem with these approaches is that the optimal size of the shape segments has to be somehow determined. Recently, Amberg *et al.* [1] proposed an approach based on local regression that fits the full SSM individually for each contour point. However, fitting the full SSM for each contour point is computationally very demanding. To deal with the computational burden, they performed the fitting only on a subset of the contour points and smoothly interpolated between the local fittings. This can also be regarded as introducing predefined segments in the model-fitting process.

In this contribution, similar to [1], the main idea is to allow a different solution in each contour-point. However, we propose a novel way to directly adapt the SSM individually for each contour-point. Smoothness constraints are introduced, ensuring a globally consistent solution. Our approach allows to obtain a locally optimal solution without the need of fitting the full SSM for each contour point and without the necessity for any predefined segments.

In section 2, we first briefly review the SSM on which we base our work. Our new approach, extending the global to a local model, is proposed in section 3, and we thoroughly evaluate the performance of our new approach in section 4.

2 Statistical Shape Model

A common and simple way to represent the shapes that are used to build the SSM is to define a set of points on every training shape. However, this requires the determination of dense point correspondences across the training shapes. The manual determination of point correspondences is a tedious and time-consuming task, which is nearly infeasible for three-dimensional models. Also, the automatic definition of point correspondences is a very demanding task that is not easy to solve [9]. For these reasons, we think that the level set representation introduced in [11] is more flexible since it does not require any point correspondences.

The original ideas for the incorporated SSM were taken from [18]. A detailed description can be found in [10]. Building the SSM starts with rigidly aligning a given set of training shapes $\{C_1, C_2, \dots, C_n\}$ with regard to translation, rotation, and scaling. Each d -dimensional shape C_i is thereby defined as the zero level-set

of a higher-dimensional function $\Phi_i : \Omega \rightarrow \mathbb{R}$, with $\Omega \subset \mathbb{N}^d$ being the input data domain. A mean level-set Φ_{mean} is obtained by averaging over all shapes.

To reduce the dimensionality of the SSM, a *Principal Component Analysis* (PCA) is performed on the set of mean-subtracted level-set functions. The PCA delivers the $n - 1$ principal modes of shape variation $\{\tilde{\Phi}_1, \dots, \tilde{\Phi}_{n-1}\}$ and corresponding variances $\{\sigma_1^2, \dots, \sigma_{n-1}^2\}$. Incorporating this information, the SSM is finally constructed by

$$\Phi_{\mathbf{w}}(\mathbf{x}) = \Phi_{\text{mean}}(\mathbf{x}) + \langle \mathbf{w}, \mathbf{b}(\mathbf{x}) \rangle . \quad (1)$$

Thereby, $\mathbf{b}(\mathbf{x}) = (\tilde{\Phi}_1(\mathbf{x}), \dots, \tilde{\Phi}_m(\mathbf{x}))^T$ is a vector of the $m \leq n - 1$ eigenshapes $\tilde{\Phi}_i$ that have most influence on shape variation, and $\mathbf{w} = (w_1, \dots, w_m)^T$ is a vector of weighting factors. The variances of the SSM are combined in the vector $\mathbf{s} = (\sigma_1, \dots, \sigma_m)^T$. All shapes $C_{\mathbf{w}}$, represented by the SSM, can be obtained through adapting the weights \mathbf{w} . They are given as the zero level-set of $\Phi_{\mathbf{w}}$.

3 Locally Deformable Statistical Shape Model

Modifying the weight-vector \mathbf{w} in eq. (1) gives us the ability to deform the whole SSM in a way predetermined by the training data. In a partitioned SSM, the idea is to increase the model flexibility by using one weight vector for each partition and fitting the model only to local regions of the data under analysis. A smooth crossover between adjacent partitions has to be ensured in order to obtain a continuous segmentation result.

To avoid the need of predefined partitions, we propose to further extend the partitioned models by allowing one weight-vector $\mathbf{v}(\mathbf{x})$ in each coordinate \mathbf{x} of the data domain Ω . A smooth transition between neighboring weights thereby guarantees a continuous segmentation result that is in consistence with the SSM in local areas. Hence, our *Locally Deformable SSM* (LDSSM) is defined as

$$\Phi_{\mathbf{v}}(\mathbf{x}) = \Phi_{\text{mean}}(\mathbf{x}) + \langle \mathbf{v}(\mathbf{x}), \mathbf{b}(\mathbf{x}) \rangle , \quad (2)$$

with $\mathbf{v} : \Omega \rightarrow \mathbb{R}^m$ being a sufficient smooth field of weight-vectors. According to eq. (1), local deformation is achieved by modifying the field of weight-vectors \mathbf{v} . All shapes $C_{\mathbf{v}}$, represented by the LDSSM, are given as the zero level-set of $\Phi_{\mathbf{v}}$.

Deformation Framework. Now that we have defined our LDSSM, we need to describe a framework that enables us to deform the LDSSM in order to approximate a given target function Φ_{targ} . For this purpose, we assume that an initial field of weight vectors \mathbf{v}_{init} exists. It may be obtained by fitting the SSM in eq. (1) to Φ_{targ} , yielding a globally optimal solution \mathbf{w}_{opt} , and then setting $\mathbf{v}_{\text{init}}(\mathbf{x}) = \mathbf{w}_{\text{opt}}$ for all \mathbf{x} .

Neglecting the smoothness constraint of our LDSSM, a weight-vector field \mathbf{v} can always be determined in a way such that the distance between the LDSSM and the target is zero by solving $\Phi_{\mathbf{v}}(\mathbf{x}) = \Phi_{\text{targ}}(\mathbf{x})$ independently for each \mathbf{x} . This is an under-determined equation system with an infinite number of solutions.

The solution we are interested in is a minimum of the corresponding error function $r(\mathbf{v}(\mathbf{x})) = (\Phi_{\mathbf{v}}(\mathbf{x}) - \Phi_{\text{targ}}(\mathbf{x}))^2$ that resides close to the initial solution $\mathbf{v}_{\text{init}}(\mathbf{x})$. This is a convex optimization problem. Hence, every local minimum is also a global minimum. We obtain the desired solution by using Cauchy's steepest descent method. Starting from $\mathbf{v}^0(\mathbf{x}) = \mathbf{v}_{\text{init}}(\mathbf{x})$, the LDSSM is evolved towards the desired solution Φ_{targ} by

$$\mathbf{v}^{k+1}(\mathbf{x}) = \mathbf{v}^k(\mathbf{x}) - \alpha(\mathbf{v}^k(\mathbf{x})) \cdot \nabla r(\mathbf{v}^k(\mathbf{x})) , \quad (3)$$

where $\alpha(\mathbf{v}^k(\mathbf{x}))$ denotes the Cauchy step-size. For our quadratic optimization problem it results to $\alpha(\mathbf{v}^k(\mathbf{x})) = \frac{\nabla r(\mathbf{v}^k(\mathbf{x}))^T \cdot \nabla r(\mathbf{v}^k(\mathbf{x}))}{2 \cdot \nabla r(\mathbf{v}^k(\mathbf{x}))^T \cdot \mathbf{b}(\mathbf{x}) \cdot \mathbf{b}(\mathbf{x})^T \cdot \nabla r(\mathbf{v}^k(\mathbf{x}))}$ (For a detailed description refer to e.g. [2], chapter 18.2.5.1). The evolved shape in each iteration k is then given by $C_{\mathbf{v}^k} = \{\mathbf{x} \mid \Phi_{\mathbf{v}^k}(\mathbf{x}) = 0\}$.

Please remind that the solution obtained from iterating eq. (3) is so far *not* restricted to the subspace of feasible shapes at all. Hence, we need to extend the steepest descent in eq. (3) in a way such that it incorporates the trained shape distribution. Therefore, the original result of eq. (3) is constrained by two consecutive regularization steps. At first it is truncated by the vector of standard deviations \mathbf{s} , obtained from PCA in section 2. Secondly, the result thus obtained is averaged in a local neighborhood with radius a . Finally, we obtain our regularized gradient descent step as

$$\mathbf{v}^{k+1}(\mathbf{x}) = \frac{1}{|A(\mathbf{x})|} \sum_{\mathbf{y} \in A(\mathbf{x})} \max \left\{ \min \left\{ \tilde{\mathbf{v}}^{k+1}(\mathbf{y}), 3 \cdot \mathbf{s} \right\}, -3 \cdot \mathbf{s} \right\}, \quad (4)$$

with $\tilde{\mathbf{v}}^{k+1}(\mathbf{y}) = \mathbf{v}^k(\mathbf{y}) - \alpha(\mathbf{v}^k(\mathbf{y})) \cdot \nabla r(\mathbf{v}^k(\mathbf{y}))$ and $A(\mathbf{x}) = \{\mathbf{z} \mid \|\mathbf{x} - \mathbf{z}\| < a\}$. The operators max and min thereby represent the component-wise maximum or minimum of the incorporated vectors. Truncating the result ensures that each local fitting always remains in the subspace of feasible shapes, and averaging implicitly ensures a smooth transition of shape-weights to satisfy our smoothness condition. Hence, by iterating equation (4), we obtain a solution that is consistent with the SSM in local areas. When we reduce the radius a of the neighborhood, we can achieve a seamless transition from a globally to a locally optimal solution.

Target Function Estimation. In order to use our model for segmenting unknown data I , we need an approximation $\hat{\Phi}_{\text{targ}}$ of the target level-set function Φ_{targ} . Our solution to this problem is to obtain a local estimate $\hat{\Phi}_{\text{targ}}^k$ of the target function in a narrow band $\text{NB}(C_{\mathbf{v}^k})$ around the current segmentation result $C_{\mathbf{v}^k}$ and set $\hat{\Phi}_{\text{targ}}^k = \Phi_{\mathbf{v}^k}$ in the rest of the data domain. Thus, in each iteration k , the weight-vectors outside the narrow band are only effected by truncation and averaging, but not by gradient descent. Consequently, weight changes around the current contour are smoothly propagated over the whole data domain to obtain a globally consistent segmentation result.

We seek a target function that attracts the evolving curve to strong edges of the input data I . Therefore, we first enhance the edges by calculating $|\nabla I|$ and slightly smoothing the result with a Gaussian kernel G_σ .

Next, we compute the gradient of the smoothed edge volume $\mathbf{g} = \nabla[G_\sigma * |\nabla I|]$ to obtain a vector field directed towards the edges in the input data.

The estimation of the target function $\hat{\Phi}_{\text{targ}}^k$ can then be modeled as a transport problem $\frac{\partial}{\partial t}\Phi(\mathbf{x}, t) = \langle \mathbf{g}(\mathbf{x}), \nabla\Phi(\mathbf{x}, t) \rangle$, with velocity $-\mathbf{g}$ and initial value $\Phi(\mathbf{x}, 0) = \Phi_{\mathbf{v}^k}(\mathbf{x})$. To solve this transport problem, we apply Euler's forward-discretization method: $\Phi^{n+1}(\mathbf{x}) = \Phi^n(\mathbf{x}) + h \cdot \langle \mathbf{g}(\mathbf{x}), \nabla\Phi^n(\mathbf{x}) \rangle$, with $\Phi^0(\mathbf{x}) = \Phi(\mathbf{x}, 0)$. Using the empirically determined step-size

$$h = \left[\max_{\mathbf{x} \in \text{NB}(C_{\mathbf{v}^k})} |\langle \mathbf{g}(\mathbf{x}), \nabla\Phi_{\mathbf{v}^k}(\mathbf{x}) \rangle| \right]^{-1}, \quad (5)$$

already one Euler-step is sufficient for an adequate approximation of the target function $\hat{\Phi}_{\text{targ}}^k$:

$$\hat{\Phi}_{\text{targ}}^k(\mathbf{x}) = \Phi^1(\mathbf{x}) = \Phi_{\mathbf{v}^k}(\mathbf{x}) + h \cdot \langle \mathbf{g}(\mathbf{x}), \nabla\Phi_{\mathbf{v}^k}(\mathbf{x}) \rangle. \quad (6)$$

Eq. (6) has the desirable property that the current segmentation result $C_{\mathbf{v}^k}$ is mostly attracted by edges parallel to the current contour.

4 Comparison of Global and Local Model

We compare the segmentation capabilities of the SSM with our proposed LDSSM by segmenting the outer paranasal sinus boundary. The knowledge of this boundary is, e.g., important in robot assisted sinus surgery [15]. 49 two-dimensional CT slices from different patients are used for the evaluation. However, it should be noted that the LDSSM is suited for three-dimensional segmentation tasks as well. The evaluation is performed in a leave-one-out fashion, meaning that we build the SSM using 48 datasets to segment the remaining dataset. The CT datasets have been rigidly aligned using anatomical landmarks, and we empirically determined that the $m = 10$ most important eigenshapes, obtained from PCA, are sufficient to represent the model-space.

Fitting the SSM. Let I be the CT slice under consideration. To segment I with the help of the SSM, we use a slightly modified version of the approach proposed in [10]. To put it briefly, the well-known Nelder-Mead simplex method, introduced in [14], is used to minimize $O_{\mathbf{w}} = |C_{\mathbf{w}}|^{-1} \sum_{\mathbf{x} \in C_{\mathbf{w}}} [G_\sigma * |\nabla I|](\mathbf{x})$, with

Table 1. Results of the error measures obtained from fitting the SSM and the LDSSM. The results were averaged over all datasets, and the significance was evaluated using the two-sided Wilcoxon rank sum test. The resulting p -values are given in the last row.

	e_{rms} : RMSE	e_h : Hausdorff distance	e_j : Dice coefficient	e_d : Jaccard coefficient
SSM	3.39 mm	11.33 mm	0.170	0.094
LDSSM	1.83 mm	7.70 mm	0.094	0.050
p -value	$3.61 \cdot 10^{-9}$	$1.83 \cdot 10^{-4}$	$8.58 \cdot 10^{-13}$	$8.59 \cdot 10^{-13}$

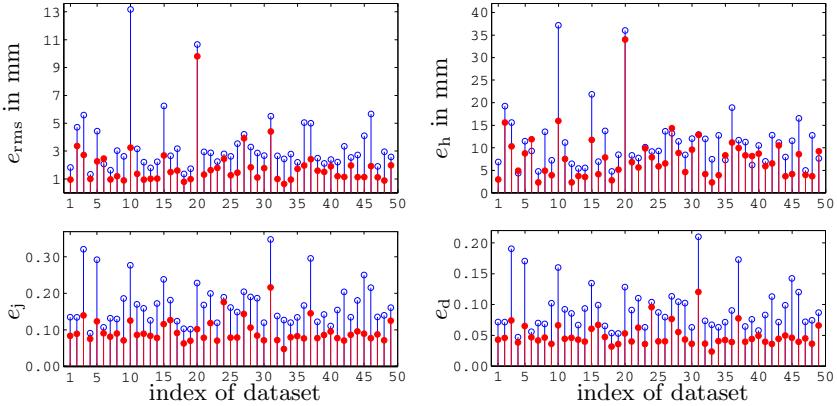


Fig. 1. Results of the error measures obtained from fitting the SSM (empty circles) and the LDSSM (filled circles), respectively. The results are plotted against each dataset.

respect to \mathbf{w} , in order to obtain the global fitting solution $\mathbf{w}_{\text{opt}} = \arg \min O_{\mathbf{w}}$. This objective function is minimal when the zero-level set $C_{\mathbf{w}}$ of the SSM captures high image gradients of the input slice I. Please note that $O_{\mathbf{w}}$ uses the same input information as the target function proposed in (6) so that a direct comparison of the global and local fitting result is possible.

Fitting the LDSSM. The local fitting is performed as described in section 3, with $\mathbf{v}_{\text{init}}(\mathbf{x}) = \mathbf{w}_{\text{opt}}$ for all \mathbf{x} and $a = 2.3$ mm. We use a slightly modified version of the target function $\hat{\Phi}_{\text{targ}}^k$ to obtain a faster convergence. The first modification is that we always force $C_{\mathbf{v}^k}$ outwards when it is located in an air-filled region, because it is very likely that an air-filled region belongs to the paranasal sinuses. Additionally, we always force the curve inwards when it is located in a bony region. The modifications are realized by adding an offset γ to eq. (6), where $\gamma = -0.1$ when $I(\mathbf{x}) < -200$, $\gamma = 0.1$ when $I(\mathbf{x}) > 200$, and $\gamma = 0$ elsewhere.

Evaluation. To evaluate the local and the global segmentation results, we calculate four error measures. These include the root mean squared error e_{rms} and the Hausdorff distance e_h (see [10] for a definition). In addition, we calculate the Jaccard coefficient $e_j = 1 - \frac{|A_{\text{res}} \cap A_{\text{ref}}|}{|A_{\text{res}} \cup A_{\text{ref}}|}$ and the Dice coefficient $e_d = 1 - \frac{2 \cdot |A_{\text{res}} \cap A_{\text{ref}}|}{|A_{\text{res}}| + |A_{\text{ref}}|}$. A_{ref} and A_{res} denote the regions enclosed by a hand-segmented reference contour and the segmentation result, respectively.

As can be seen in table 1 and in fig. 1, our proposed LDSSM is able to deliver more accurate segmentation results compared to the SSM. The statistical significance of this finding was confirmed using the two-sided Wilcoxon rank sum test [8] (see table 1 for the corresponding ρ -values). Especially for distinctive frontal sinuses the LDSSM is clearly better suited. This can be seen in fig. 2.

However, there exist six datasets where the Hausdorff distance of the global fit is smaller than the corresponding local result. This is because in a very few

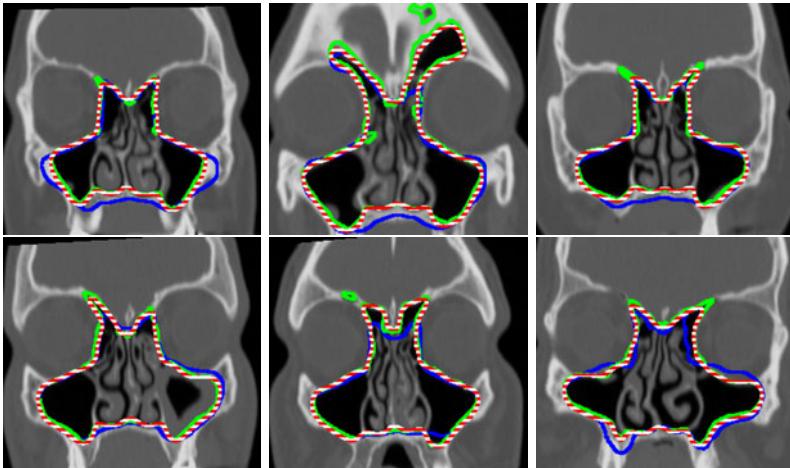


Fig. 2. Chosen CT slices, showing the hand-segmented reference (green line) and results obtained with the SSM (blue line) and the LDSSM (dashed line). It is clearly visible that the LDSSM is able to segment the nasal cavities more precisely than the SSM.

local regions the hand-segmented reference does not correspond to any edge of the input data, leading to a conflict between our target definition in eq. (6) and the hand-segmented reference. Consequently, the LDSSM adapts to a nearby edge that can be expressed through linear combinations of the training shapes.

5 Summary and Outlook

We presented a *Locally Deformable Statistical Shape Model* (LDSSM) that allows for a local fit of the corresponding SSM in each contour point. The proposed LDSSM is made up of a field of weight-vectors $v(x)$, one for each coordinate x , and a smoothness condition that ensures a seamless transition between neighboring contour points. Moreover, we derived a framework that enables us to evolve the LDSSM by directly modifying the weight-vector-field. In contrast to previously proposed partitioned shape models, we do not need predefined segments.

Experimental results showed the great potential for using the LDSSM in segmentation tasks. We were able to improve the initial solution, obtained by fitting the SSM, in almost all datasets under consideration. Inspired by Cootes *et al.* [6], we further plan to incorporate statistical information about the intensity values on the training shapes into our LDSSM. This should help us to refine our segmentation results in regions where no distinctive edge information is present.

References

1. Amberg, M., Lüthi, M., Vetter, T.: Local regression based statistical model fitting. In: Goesele, M., Roth, S., Kuijper, A., Schiele, B., Schindler, K. (eds.) DAGM 2010. LNCS, vol. 6376, pp. 452–461. Springer, Heidelberg (2010)

2. Bronshtein, I., Semendyayev, K., Musiol, G., Mühlig, H.: *Handbook of Mathematics*, 5th edn. Springer, New York (2007)
3. de Bruijne, M., van Ginneken, B., Viergever, M.A., Niessen, W.J.: Adapting active shape models for 3D segmentation of tubular structures in medical images. In: Taylor, C.J., Noble, J.A. (eds.) *IPMI 2003. LNCS*, vol. 2732, pp. 136–147. Springer, Heidelberg (2003)
4. Cootes, T., Taylor, C.: Combining point distribution models with shape models based on finite element analysis. *Image and Vision Comput.* 13(5), 403–409 (1995)
5. Cootes, T., Taylor, C.: Data driven refinement of active shape model search. In: *Proceedings of the British Machine Vision Conference*, pp. 383–392 (1996)
6. Cootes, T., Taylor, C., Cooper, D., Graham, J.: Active shape models - their training and application. *Computer Vision and Image Understanding* 61(1), 38–59 (1995)
7. Davatzikos, C., Tao, X., Shen, D.: Hierarchical active shape models, using the wavelet transform. *IEEE Transactions on Medical Imaging* 22(3), 414–423 (2003)
8. Gibbons, J., Chakraborti, S.: *Nonparametric Statistical Inference*, 5th edn. CRC Press, Boca Raton (2010)
9. Heimann, T., Meinzer, H.: Statistical shape models for 3d medical image segmentation: A review. *Medical Image Analysis* 13(4), 543–563 (2009)
10. Last, C., Winkelbach, S., Wahl, F., Eichhorn, K., Bootz, F.: A model-based approach to the segmentation of nasal cavity and paranasal sinus boundaries. In: Goesele, M., Roth, S., Kuijper, A., Schiele, B., Schindler, K. (eds.) *DAGM 2010. LNCS*, vol. 6376, pp. 333–342. Springer, Heidelberg (2010)
11. Leventon, M., Grimson, W., Faugeras, O.: Statistical shape influence in geodesic active contours. In: *Proc. of CVPR*. pp. 316–323 (2000)
12. Lötjönen, J., Antila, K., Lamminmäki, E., Koikkalainen, J., Lilja, M., Cootes, T.F.: Artificial enlargement of a training set for statistical shape models: Application to cardiac images. In: Frangi, A., Radeva, P., Santos, A., Hernandez, M. (eds.) *FIMH 2005. LNCS*, vol. 3504, pp. 92–101. Springer, Heidelberg (2005)
13. Nain, D., Haker, S., Bobick, A., Tannenbaum, A.: Multiscale 3-d shape representation and segmentation using spherical wavelets. *IEEE Transactions on Medical Imaging* 26(4), 598–618 (2007)
14. Nelder, J., Mead, R.: A simplex method for function minimization. *The Computer Journal* 7(4), 308–313 (1965)
15. Rilk, M., Wahl, F., Eichhorn, K., Wagner, I., Bootz, F.: Path planning for robot-guided endoscopes in deformable environments. In: Kröger, T., Wahl, F. (eds.) *Advances in Robotics Research*, pp. 263–274. Springer, Heidelberg (2009)
16. Shang, Y., Dössel, O.: Statistical 3d shape-model guided segmentation of cardiac images. In: *Computers in Cardiology*, vol. 31, pp. 553–556 (2004)
17. Shen, D., Herskovits, E., Davatzikos, C.: An adaptive-focus statistical shape model for segmentation and shape modeling of 3-d brain structures. *IEEE Transactions on Medical Imaging* 20(4), 257–270 (2001)
18. Tsai, A., Yezzi Jr., A., Wells, W., Tempany, C., Tucker, D., Fan, A., Grimson, W., Willsky, A.: A shape-based approach to the segmentation of medical imagery using level sets. *IEEE Transactions on Medical Imaging* 22(2), 137–154 (2003)
19. Weese, J., Kaus, M., Lorenz, C., Lobregt, S., Truyen, R., Pekar, V.: Shape constrained deformable models for 3D medical image segmentation. In: Insana, M., Leahy, R. (eds.) *IPMI 2001. LNCS*, vol. 2082, pp. 380–387. Springer, Heidelberg (2001)
20. Zhao, Z., Aylward, S., Teoh, E.: A novel 3D partitioned active shape model for segmentation of brain MR images. In: Duncan, J., Gerig, G. (eds.) *MICCAI 2005. LNCS*, vol. 3749, pp. 221–228. Springer, Heidelberg (2005)

Monte Carlo Expectation Maximization with Hidden Markov Models to Detect Functional Networks in Resting-State fMRI

Wei Liu¹, Suyash P. Awate¹, Jeffrey S. Anderson², Deborah Yurgelun-Todd³, and P. Thomas Fletcher¹

¹ Scientific Computing and Imaging Institute, University of Utah, USA

² Department of Radiology, University of Utah, USA

³ Department of Psychiatry, University of Utah, USA

Abstract. We propose a novel Bayesian framework for partitioning the cortex into distinct functional networks based on resting-state fMRI. Spatial coherence within the network clusters is modeled using a hidden Markov random field prior. The normalized time-series data, which lie on a high-dimensional sphere, are modeled with a mixture of von Mises-Fisher distributions. To estimate the parameters of this model, we maximize the posterior using a Monte Carlo expectation maximization (MCEM) algorithm in which the intractable expectation over all possible labelings is approximated using Monte Carlo integration. We show that MCEM solutions on synthetic data are superior to those computed using a mode approximation of the expectation step. Finally, we demonstrate on real fMRI data that our method is able to identify visual, motor, salience, and default mode networks with considerable consistency between subjects.

1 Introduction

Resting-state functional magnetic resonance imaging (fMRI) measures background fluctuations in the blood oxygen level-dependent (BOLD) signal of the brain at rest. The temporal correlations between these signals are used to estimate the functional connectivity of different brain regions. This technique has shown promise as a clinical research tool to describe functional abnormalities in Alzheimer's disease, schizophrenia, and autism [4]. In resting-state fMRI, a standard analysis procedure is to select a region of interest (ROI), or seed region, and find the correlation between the average signal of the ROI and other voxels in the brain. These correlations are thresholded so that only those voxels with significant correlations with the seed region are shown. Recent methods to find functional networks without seed regions include independent component analysis (ICA) [2], which often includes an ad hoc method to manually choose those components that are anatomically meaningful. Other approaches employ clustering techniques to automatically partition the brain into functional networks. A similarity metric is defined, e.g., correlation [5] or frequency coherence [9], and then a clustering method such as k -means or spectral clustering is used to group

voxels with similar time series. A drawback of these approaches is that they disregard the spatial position of voxels, and thus ignore the fact that functional networks are organized into sets of spatially coherent regions.

We introduce a new data-driven method to partition the brain into networks of functionally-related regions from resting-state fMRI. The proposed algorithm does not require specification of a seed, and there is no ad hoc thresholding or parameter selection. We make a natural assumption that functionally homogeneous regions should be spatially coherent. Our method incorporates spatial information through a Markov random field (MRF) prior on voxel labels, which models the tendency of spatially-nearby voxels to be within the same functional network. Each time series is first normalized to zero mean and unit norm, which results in data lying on a high-dimensional unit sphere. We then model the normalized time-series data as a mixture of von Mises-Fisher (vMF) distributions [1]. Each component of the mixture model corresponds to the distribution of time series from one functional network. Solving for the parameters in this combinatorial model is intractable, and we therefore use a stochastic method called Monte Carlo Expectation Maximization (MCEM), which approximates the expectation step using Monte Carlo integration. The stochastic property of MCEM makes it possible to explore a large solution space, and it performs better than a standard mode approximation method using iterated conditional modes (ICM).

The proposed method in this paper is related to previous approaches using MRFs to model spatial relationships in fMRI data. Descombes et al. [3] use a spatio-temporal MRF to analyze task-activation fMRI data. Liu et al. [7] use an MRF model of resting state fMRI to estimate pairwise voxel connections. However, neither of these approaches tackle the problem of clustering resting-state fMRI into functional networks.

2 Hidden Markov Models of Functional Networks

We use a Bayesian statistical framework to identify functional networks of the gray matter in fMRI data. We formulate a generative model, which first generates a spatial configuration of functional networks in the brain, followed by an fMRI time series for each voxel based on its network membership. We employ an MRF prior to model network configurations, represented via unknown, or *hidden*, labels. Given a label, we assume that the fMRI time series, normalized to zero mean and unit norm, are drawn from a von Mises-Fisher likelihood.

Let $\mathcal{S} = \{1, \dots, N\}$ be the set of indices for all gray-matter voxels. We assume that the number of networks L is a known free parameter. Let $\mathcal{L} = \{1, 2, \dots, L\}$ be the set of labels, one for each network. We denote a label map for functionally-connected networks as a vector $\mathbf{z} = (z_1, \dots, z_N)$, $z_i \in \mathcal{L}$. Let $\mathcal{Z} = \mathcal{L}^N$ be the set of all possible \mathbf{z} configurations.

2.1 Markov Prior Model

Functional networks should consist of few, reasonably-sized, possibly distant regions. We model such networks \mathbf{z} using the *Potts* MRF [6]:

$$P(\mathbf{z}) = \frac{1}{C} \exp \left(-\beta \sum_{i \in \mathcal{S}} \sum_{j \in \mathcal{N}_i} T(z_i \neq z_j) \right),$$

where T is 1 if its argument is true and 0 otherwise; \mathcal{N}_i is the set of neighbors of i as defined by the neighborhood system underlying the MRF; $\beta > 0$ is a model parameter controlling label-map smoothness; C is a normalization constant that is the sum of $P(\mathbf{z})$ over all possible configuration of \mathbf{z} . The Markov-Gibbs equivalence [6] implies that the conditional distribution of z_i at site i is:

$$P(z_i | \mathbf{z}_{-i}) = P(z_i | z_{\mathcal{N}_i}) = \frac{\exp \left(-\beta \sum_{j \in \mathcal{N}_i} T(z_i \neq z_j) \right)}{\sum_{l \in \mathcal{L}} \exp \left(-\beta \sum_{j \in \mathcal{N}_i} T(l \neq z_j) \right)}, \quad (1)$$

where \mathbf{z}_{-i} is the collection of all variables in \mathbf{z} excluding site i . The neighborhood is the usual 6 adjacent voxels, which does not overly smooth across boundaries. Previous works [10,3] have demonstrated the advantages of MRF's over Gaussian smoothing in preserving segment boundaries.

2.2 Likelihood Model

To make the analysis robust to shifts or scalings of the data, we normalize the time series at each voxel to zero mean and unit length. This results in the data being projected onto a high-dimensional unit sphere. After normalization, the sample correlation between two time series is equal to their inner product, or equivalently, the cosine of the geodesic distance between these two points on the sphere. Thus, we re-formulate the problem of finding clusters of voxels with high correlations to the problem of finding clusters with small within-cluster distances on the sphere.

We use the notation $\mathbf{x} = \{(\mathbf{x}_1, \dots, \mathbf{x}_N) | \mathbf{x}_i \in S^{p-1}\}$ to denote the set of *normalized* time series. Observe that given $\mathbf{z} \in \mathcal{Z}$, the random vectors \mathbf{x}_i are conditional independent. Thus, the likelihood $\log P(\mathbf{x} | \mathbf{z}) = \sum_{i \in \mathcal{S}} \log P(\mathbf{x}_i | z_i)$. We model the emission function $P(\mathbf{x}_i | z_i)$ using the von Mises-Fisher distribution

$$f(\mathbf{x}_i; \boldsymbol{\mu}_l, \kappa_l | z_i = l) = C_p(\kappa_l) \exp (\kappa_l \boldsymbol{\mu}_l^T \mathbf{x}_i), \quad \mathbf{x}_i \in S^{p-1}, \quad l \in \mathcal{L} \quad (2)$$

where, for the cluster labeled l , $\boldsymbol{\mu}_l$ is the mean direction, $\kappa_l \geq 0$ is the *concentration parameter*, and the normalization constant $C_p(\kappa) = \kappa^{\frac{p}{2}-1} / \{(2\pi)^{\frac{p}{2}} I_{\frac{p}{2}-1}(\kappa)\}$, where I_ν denotes the modified Bessel function of the first kind with order ν . The larger the κ , the greater is the density concentrated around the mean direction. Since (2) depends on \mathbf{x} only by $\boldsymbol{\mu}^T \mathbf{x}$, the vMF distribution is unimodal and rotationally symmetric around $\boldsymbol{\mu}$.

In the Bayesian framework, we also define distributions on parameters. We assume that $\forall l \in \mathcal{L}, \kappa_l \sim \mathcal{N}(\mu_\kappa, \sigma_\kappa^2)$ with hyperparameters μ_κ and σ_κ^2 that can be set empirically. This prior enforces constraints that the clusters should not have extremely high or low concentration parameters. We empirically tune the hyperparameters μ_κ and σ_κ^2 and have found the results to be robust to specific choices of the hyperparameters.

3 Monte Carlo EM

To estimate the model parameters and the hidden labels, we use a stochastic variant of expectation maximization (EM) called Monte Carlo EM (MCEM) [10]. The standard EM algorithm maximizes the expectation of the log-likelihood of joint pdf of \mathbf{x} and the hidden variable \mathbf{z} with respect to the posterior probability $P(\mathbf{z}|\mathbf{x})$, i.e. $\mathbb{E}_{P(\mathbf{z}|\mathbf{x})}[\log P(\mathbf{x}, \mathbf{z}; \boldsymbol{\theta})]$. The combinatorial number of configurations for \mathbf{z} makes this expectation intractable. Thus, we use Monte Carlo simulation to approximate this expectation as

$$\tilde{Q}(\boldsymbol{\theta}; \mathbf{x}, \mathbf{z}) \approx \frac{1}{M} \sum_{m=1}^M \log P(\mathbf{z}^m; \beta) + \log P(\mathbf{x}|\mathbf{z}^m; \boldsymbol{\theta}_L), \quad (3)$$

where \mathbf{z}^m is a sample from $P(\mathbf{z}|\mathbf{x})$, $\boldsymbol{\theta}_L = \{\boldsymbol{\mu}_l, \kappa_l : l \in \mathcal{L}\}$ is the parameter vector of the likelihood, and $\boldsymbol{\theta} = \{\beta, \boldsymbol{\theta}_L\}$ is the full parameter vector of the model. Computing the MRF prior in (3) is still intractable due to the normalization constant, and we instead use a pseudo-likelihood approximation [6], which gives

$$\tilde{Q} \approx \frac{1}{M} \sum_{m=1}^M \sum_{i \in \mathcal{S}} \log P(z_i|z_{\mathcal{N}_i}; \beta) + \frac{1}{M} \sum_{m=1}^M \sum_{i \in \mathcal{S}} \log P(\mathbf{x}_i|z_i; \boldsymbol{\theta}_L) = \tilde{Q}_P + \tilde{Q}_L.$$

We use \tilde{Q}_P to denote the log-pseudo-likelihood of the prior distribution, and use \tilde{Q}_L to denote the log-likelihood distribution.

3.1 Sampling from the Posterior

Given the observed data \mathbf{x} and parameter value $\boldsymbol{\theta} = \{\beta, \boldsymbol{\theta}_L\}$, we sample from the posterior distribution $P(\mathbf{z}|\mathbf{x}; \boldsymbol{\theta})$ using Metropolis sampling. We define the posterior energy, which is to be minimized, as the negative log of the posterior $P(z_i|\mathbf{x}_i)$. Thus, Bayes rule implies:

$$U(z_i = l|\mathbf{x}) = \beta \sum_{j \in \mathcal{N}_i} T(z_i \neq z_j) - \log C_p(\kappa_l) - \kappa_l \boldsymbol{\mu}_l^T \mathbf{x}_i + \text{const}, \quad (4)$$

which is the sum of the prior energy, the conditional energy, and a parameter-independent quantity. Then, given a current configuration \mathbf{z}^n Metropolis sampling generates a new candidate label map \mathbf{w} as follows: (i) draw a new label l' at site i with uniform distribution; \mathbf{w} has value l' at site i , with other sites remaining the same as \mathbf{z}^n ; (ii) compute the change of energy $\Delta U(\mathbf{w}) = U(\mathbf{w}|\mathbf{x}) - U(\mathbf{z}^n|\mathbf{x}) = U(z_i = l'|\mathbf{x}) - U(z_i = l|\mathbf{x})$; (iii) accept candidate \mathbf{w} as \mathbf{z}^{n+1} with probability $\min(1, \exp\{-\Delta U(\mathbf{w})\})$; (iv) after a sufficiently long burn-in period, generate a sample of size M from the posterior distribution $P(\mathbf{z}|\mathbf{x})$.

3.2 Parameter Estimation

Estimating $\boldsymbol{\theta}_L$: By maximizing \tilde{Q}_L with the constraint $\|\boldsymbol{\mu}_l\| = 1$, we get

$$R_l = \sum_{m=1}^M \sum_{i \in \mathcal{S}_l} \mathbf{x}_i, \quad \hat{\boldsymbol{\mu}}_l = \frac{R_l}{\|R_l\|}, \quad (5)$$

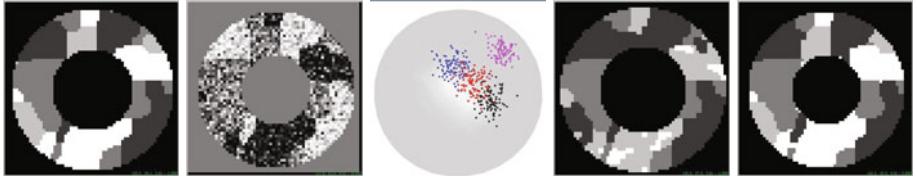


Fig. 1. Synthetic example. From left to right: true labels, first time point of observed time series, time series plot on sphere, label map estimated by mode-approximation, and label map estimated by MCEM.

where $S_l = \{i \in \mathcal{S} : z_i = l\}$ is the set of data points in cluster l . We have no *a priori* knowledge for μ_l , so a maximum likelihood estimation in (5) is the best we can do. For κ_l we maximize the posterior distribution $P(\kappa_l | \mathbf{x}, \mathbf{z}^1, \dots, \mathbf{z}^M)$. Since \tilde{Q}_L is not dependent on κ , we maximize $\tilde{Q}_L(\kappa_l) + \log P(\kappa_l; \mu_\kappa, \sigma_\kappa^2)$ and get

$$A_p(\hat{\kappa}_l) + \frac{\hat{\kappa}_l - \mu_\kappa}{N_l \sigma_\kappa^2} = R_l, \quad (6)$$

where $A_p(\hat{\kappa}_l) = I_{\frac{p}{2}}(\hat{\kappa}_l)/I_{\frac{p}{2}-1}(\hat{\kappa}_l)$ and $N_l = |\mathcal{S}_l|$ is the number of data points in cluster l . Because (6) contains the ratio of two modified Bessel functions, an analytic solution is unavailable and we have to resort to a numerical solution. We use Newton's method for solving $g(\hat{\kappa}_l) = A_p(\hat{\kappa}_l) - (\hat{\kappa}_l - \mu_\kappa)/(N_l \sigma_\kappa^2) - R_l = 0$. The choice of initial value for Newton's algorithm depends on the strength of the prior on κ_l (i.e. the σ_κ value). For a noninformative prior, $\hat{\kappa}_l = (pR_l - R^3)/(1 - R^2)$ is a good initial value [1]. For a strong prior, a reasonable initial value is the current value of κ_l .

Estimating β : To estimate β , we again rely on Newton's method to find the solution numerically. The derivatives $\partial \tilde{Q}_P / \partial \beta$ and $\partial^2 \tilde{Q}_P / \partial \beta^2$ for the pseudo-likelihood approximation of the MRF prior are easily computed.

3.3 MCEM-Based Algorithm for Hidden-MRF Model Estimation

Given the methods for sampling and parameter estimation, we estimated the hidden-MRF model by iteratively using (i) MCEM to learn model parameters and (ii) using ICM to compute optimal network labels. In the expectation (E) step, we draw samples from the posterior $P(\mathbf{z}|\mathbf{x})$, given current estimates for parameters θ . In the maximization (M) step, we use these samples to update estimates for the parameters θ .

4 Results and Conclusion

Synthetic example: We first simulate low-dimensional time series (2D 64×64 image domain; 3 timepoints, for visualization on sphere S^2) to compare the (i) proposed method using MCEM with (ii) the mode-approximation approach

that replaces the E step in EM with a mode approximation. We simulate a label map by sampling from a MRF having $\beta = 2$. Given the label map, we simulate vMF samples (on the sphere S^2). Figure 1 shows that the MCEM solution is close to the ground truth, while the mode-approximation solution is stuck in a local maximum.

Resting-State fMRI: We evaluated the proposed method on real data, from healthy control subjects, in a resting-state fMRI study. BOLD EPI images (TR = 2.0 s, TE = 28 ms, 40 slices at 3 mm slice thickness, 64 x 64 matrix, 240 volumes) were acquired on a Siemens 3 Tesla Trio scanner. The data was preprocessed in SPM, including motion correction, registration to T2 and T1 structural MR images, spatial smoothing by a Gaussian filter, and masked to include only the gray-matter voxels. We used the conn software [11] to regress out signals from the ventricles and white matter, which have a high degree of physiological artifacts. A bandpass filter was used to remove frequency components below 0.01 Hz and above 0.1 Hz. We then projected the data onto the unit sphere by subtracting the mean of each time series and dividing by the magnitude of the resulting time series. We then applied the proposed method to estimate the functional network labels with the number of clusters set to $L = 8$.

Figure 2 shows the optimal label maps, produced by the proposed method for 3 of all 16 subjects in the dataset. We note that among the 8 clusters, one cluster, with the largest κ value and largest number of voxels, corresponds to background regions with weakest connectivity and is not shown in the figure. Among the clusters shown, we can identify the visual, motor, dorsal attention, executive control, salience, and default mode networks (DMN) [8]. Four networks: the visual, motor, executive control, and DMN, were robustly found across all subjects. More variability was found in the dorsal attention network (notice that it is much larger in subject 3) and salience network (notice that it is missing in subject 2). We found that changing the number of clusters, although leading to different label maps, preserves the four robust networks. For instance, we also ran the analysis with the number of clusters set to 4 or 6 (results not shown) and were able to recover the same four robust networks.

The next experiment compares our results with ICA. A standard ICA toolbox (GIFT; mialab.mrn.org) was applied on the same preprocessed data of each subject independently, which we call “Individual ICA”. We also applied standard Group ICA, using all data from the 16 subjects simultaneously. In both ICA experiments the number of components are set to 16. The component maps are converted to z score and thresholded at 1. For each method we computed an overlap map for each functional network by adding the corresponding binary

Table 1. The number of voxels with value greater than 8 in the overlapped label map

	DMN	Motor	Attention	Visual
MCEM	5043	7003	3731	5844
Individual ICA	114	167	228	134
Group ICA	3075	5314	3901	3509

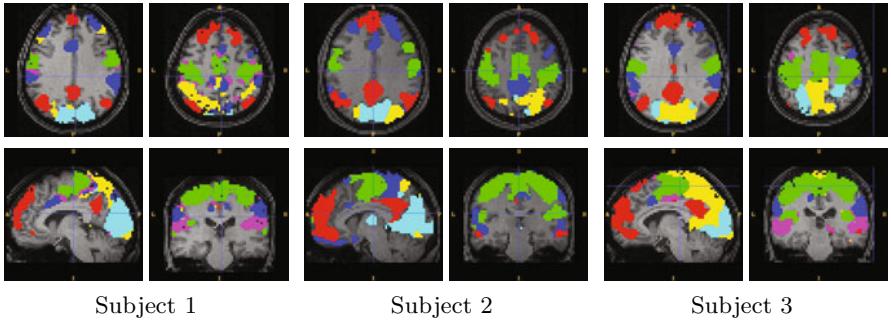


Fig. 2. Functional networks detected by the proposed method for 3 subjects overlaid on their T1 images. The clusters are the visual (cyan), motor (green), executive control (blue), salience (magenta), dorsal attention (yellow), and default mode (red) networks.

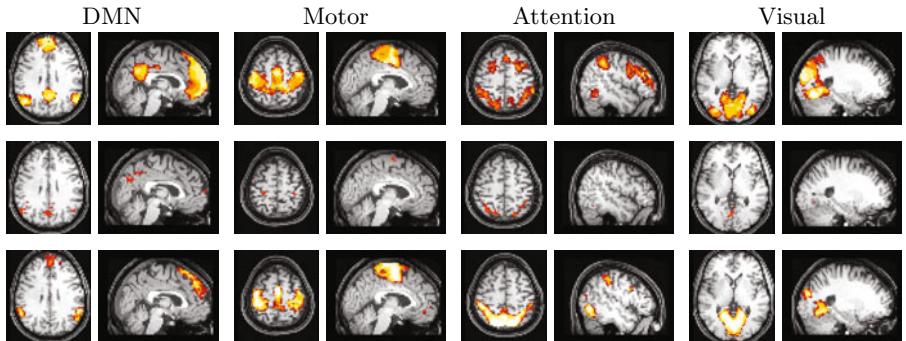


Fig. 3. Comparison of overlap label maps by our approach and ICA for 16 subjects. Top: our MCEM approach, middle: Individual ICA, bottom: Group ICA. Color ranges from 8 (red) to 16 (yellow).

label maps of all 16 subjects. The results in Figure 3 show our method can detect the motor, attention, and visual network with accuracy comparable with Group ICA. Besides, our method also detects DMN with posterior cingulate cortex (PCC) and medial prefrontal cortex (MPFC), while Group ICA split the DMN into two components, one with the MPFC and another with the PCC (not shown).

To see the consistency of the label map between subjects for all three methods, we look at each method's overlapped label map and count the number of voxels whose value are greater than 8. Table 1 shows that our method exhibits better consistency than both Individual and Group ICA.

Conclusion: We present a novel Bayesian approach to detect functional networks of the brain from resting-state fMRI that incorporates an MRF for spatial regularization, and we use MCEM to approximate the maximum posterior solution. Future work will include extending the model to group analysis, which can

be achieved using a hierarchical Bayesian model. We also plan to investigate the use of non-parametric Bayesian methods to estimate the number of clusters.

Acknowledgments. This work was funded in part by NIH Grant R01 DA020269 (Yurgelun-Todd).

References

1. Banerjee, A., Dhillon, I., Ghosh, J., Sra, S.: Clustering on the unit hypersphere using von Mises-Fisher distributions. *J. Machine Learning Res.* 6(2), 1345 (2006)
2. Beckmann, C., Smith, S.: Tensorial extensions of independent component analysis for multisubject fMRI analysis. *Neuroimage* 25(1), 294–311 (2005)
3. Descombes, X., Kruggel, F., Cramon, D.V.: Spatio-temporal fMRI analysis using Markov random fields. *IEEE Trans. on Medical Imaging* 17(6), 1028–1039 (1998)
4. Fox, M., Greicius, M.: Clinical Applications of Resting State Functional Connectivity. *Frontiers in Systems Neuroscience* 4 (2010)
5. Golland, P., Lashkari, D., Venkataraman, A.: Spatial patterns and functional profiles for discovering structure in fMRI data. In: 42nd Asilomar Conference on Signals, Systems and Computers, pp. 1402–1409 (2008)
6. Li, S.Z.: Markov random field modeling in computer vision. Springer, Heidelberg (1995)
7. Liu, W., Zhu, P., Anderson, J., Yurgelun-Todd, D., Fletcher, P.: Spatial regularization of functional connectivity using high-dimensional markov random fields. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6362, pp. 363–370. Springer, Heidelberg (2010)
8. Raichle, M.E., MacLeod, A.M., Snyder, A.Z., Powers, W.J., Gusnard, D.A., Shulman, G.L.: A default mode of brain function. *PNAS* 98(2), 676–682 (2001)
9. Thirion, B., Dodel, S., Poline, J.: Detection of signal synchronizations in resting-state fMRI datasets. *Neuroimage* 29(1), 321–327 (2006)
10. Wei, G., Tanner, M.: A Monte Carlo implementation of the EM algorithm and the poor man's data augmentation algorithms. *Journal of the American Statistical Association* 85(411), 699–704 (1990)
11. Whitfield-Gabrieli, S.: Conn Matlab toolbox (March 2011),
<http://web.mit.edu/swg/software.htm>

DCE-MRI Analysis Using Sparse Adaptive Representations

Gabriele Chiusano, Alessandra Staglianò, Curzio Basso, and Alessandro Verri

DISI, Università di Genova,
Via Dodecaneso, 35
16146 Genova, Italy

{chiusano,stagliano,basso,verri}@disi.unige.it

Abstract. Dynamic contrast-enhanced MRI (DCE-MRI) plays an important role as an imaging method for the diagnosis and evaluation of several diseases. Indeed, clinically relevant, per-voxel quantitative information may be extracted through the analysis of the enhanced MR signal. This paper presents a method for the automated analysis of DCE-MRI data that works by decomposing the enhancement curves as sparse linear combinations of elementary curves learned without supervision from the data. Experimental results show that performances in denoising and unsupervised segmentation improve over parametric methods.

1 Introduction

With advances in medical imaging, the use of Dynamic Contrast-Enhanced MRI (DCE-MRI) is becoming more and more common. This powerful imaging modality allows for the investigation of biological processes along time, that is often essential for the accurate understanding of a number of diseases.

DCE-MRI is an acquisition technique that provides information about perfusion, capillary permeability and tissue vascularity and has evolved as an important method for the evaluation of several diseases such as prostate cancer [3], breast cancer [1,14] cardiac and cerebral ischemia [8], renal dysfunctions [18] and rheumatoid arthritis [9]. The changes of signal enhancement and image intensities during and after the injection of a contrast agent (CA) reflect changes in local concentration of CA that are proportional both to the extent and severity of the disease. Clinically relevant, quantitative information is typically extracted by voxel-wise fitting of the signal enhancement curves with a parametric model and possibly defining, either manually or automatically, the regions of interests where the analysis should be carried out.

This paper presents an automated method for processing and analyzing DCE-MRI data based on dictionary learning techniques [13,11,2]. The proposed method learns a dictionary of elementary curves whose sparse linear combination can be used to accurately and effectively represent the enhancement curves. Dictionary learning techniques present a number of advantages compared to more traditional signal processing approaches, such as better fitting of the distribution of the data thanks to the adaptivity of the method [15,2], very effective denoising

capabilities [2] and state of the art discriminative power of the sparse decompositions in object detection and classification [11,17]. Moreover, assuming that each curve is the combination of elementary patterns naturally takes into account partial volume effects of MRI acquisitions. Compared to non-adaptive methods for DCE-MRI analysis, e.g. based on pharmacokinetic models [7] or analytic models [1,9], no assumptions on the curve form is made. Indeed, the method is applied to two different clinical settings, namely musculoskeletal imaging and urology imaging, in order to prove that adaptive and sparse representations can be applied effortlessly to different types of data without loosing effectiveness in discrimination tasks and, more in general, supporting them as better descriptors than non-sparse, non-adaptive ones. The method was first validated on synthetic data, then tested on real DCE-MRI studies of wrists, acquired from pediatric patients affected by juvenile idiopathic arthritis (JIA), and kidneys, acquired from pediatric patients affected by renal disfunctions (RD). The experimental results were assessed both qualitatively and quantitatively, and compared with the methods presented, respectively, in [1,7] for the synthetic and wrist datasets and in [18] for the kidneys dataset.

2 Methods

In the following, x_i^t denotes the enhancement value for the i -th pixel at time t , defined as the difference between its intensity value at time t and at time 0. The vector storing the enhancements for all d acquisition times is denoted by $x_i \in \mathbb{R}^d$, from now on referred to as enhancement curve (EC). In order to reduce computation time, the analysis of the ECs is performed on a subset of all the curves. All voxels belonging to the background or to non-enhancing areas are automatically pruned from the set of voxels to be processed. The i -th voxel is selected on the basis of the mean \bar{x}_i and standard deviations σ_i of x_i^t : if \bar{x}_i and σ_i are lower than given thresholds, the voxel is pruned. The subset of voxels to be analyzed may be further reduced if a region of interest (ROI) has been provided by the user or has been automatically defined.

2.1 Dictionary Learning on Enhancement Curves

Dictionary learning (DL) methods are based on the assumption that input data can be represented as sparse combinations of appropriately designed elementary signals, called atoms. In the present context, this is equivalent to assume that there exists a dictionary of k atoms $\{d_1, \dots, d_k\}$ such that $x_i \approx \sum_j \alpha_{ij} d_j$ for some choice of α_{ij} . In the following $A \in \mathbb{R}^{k \times n}$ denotes the matrix with coefficients α_{ij} , $\alpha_i \in \mathbb{R}^k$ its column vectors, and $D \in \mathbb{R}^{d \times k}$ the matrix whose columns are the dictionary atoms. Rather than designing the atoms up-front, dictionary learning aims at learning a good set of atoms from examples. In order to learn a dictionary from the set of ECs x_i selected from a DCE-MRI, the following problem is solved

$$\min_{D,A} \sum_{i=1}^n \|x_i - D\alpha_i\|^2 + \tau \|\alpha_i\|_1 \quad \text{subject to} \quad \|d_i\|_2 \leq 1. \quad (1)$$

The first term in the summation is a data-fidelity term that ensures a good reconstruction, while the second term has been shown to induce sparsity in the decompositions α_i . Sparsity is a desirable feature of the decompositions, because it improves their discriminative power when used in subsequent tasks. The norm of the atoms is constrained to have unitary maximal norm, in order to avoid equivalent solutions where the atoms have large norms and the decompositions are not sparse. There is not any control to avoid redundancy of the atoms, as shown in Figure 1: in case of similar atoms the weight will be distributed in a way that will not affect the sparsity. The coefficient τ weights the sparsity penalty and is a free parameter. Although the joint minimization problem in (D, A) is non-convex and non-differentiable, it has been shown that it can be solved quite efficiently by iteratively minimizing first with respect to A and then to D , a procedure known as block-coordinate descent:

1. Initialize D and A ;
2. solve problem (1) with fixed D ;
3. solve problem (1) with fixed A ;
4. go back to step (2) until convergence.

Both subproblems (2) and (3) are convex and can be solved efficiently. For a detailed description of the optimization please refer to [11,12].

To summarize, for a given choice of the parameters k and τ and of the initialization, the solution of problem (1) yields a dictionary of k elementary curve patterns, the decompositions A of the input data X , and the reconstructions DA . As shown in many papers [2,15] and confirmed in the present work, the reconstructions are denoised versions of the inputs.

2.2 Unsupervised Segmentation

As well as for denoising purposes, the representations α_i are used for the unsupervised segmentation of the input images. To this aim, the k-means algorithm is used: given a dataset of n ECs x_i and their representations α_i , the algorithm aims at partitioning the dataset in K clusters in which each observation belongs to the cluster with the nearest mean. For more details, see [4, Sec. 9.1].

3 Results

The proposed method was evaluated on simulated and on real DCE-MRI sequences. In both cases the analysis was performed voxel-by-voxel on the ECs. All experiments were implemented and run using **Python**¹ programming language, **Scipy**², a Python package for mathematics, science, and engineering, and **Scikits.learn**³, a Python package integrating classic machine learning

¹ <http://www.python.org>

² <http://www.scipy.org/>

³ <http://scikit-learn.sourceforge.net/>

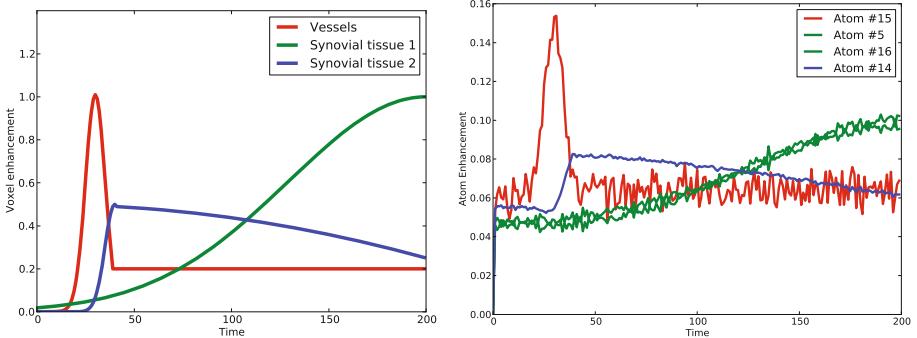


Fig. 1. Left: the three different types of enhancement curves generated, see Lavini et al.[10]. Right: the four most used atoms, corresponding to the EC patterns associated with each phantom regions. Colors match with the enhancement curve prototypes.

algorithms in Scipy. The dictionary learning problem was solved using an open-source Python implementation we developed that is available at PADDLE⁴.

In order to do a quantitative comparison over the denoising task, the performances were assessed by computing the Peak Signal to Noise Ratio (PSNR), as specified in [2]. PSNR is defined as: $PSNR = 20 \cdot \log_{10} \left(\frac{\max_i \|x_i\|_2}{\|X - DA\|_F^2 / n} \right)$. Furthermore, the proposed method was compared with two other approaches from Guo et al. [7] and Agner et al.[1] over the synthetic and wrists dataset, and with Zöellner et al. [18] over the kidneys dataset. All these works are good representatives of methods found in the literature.

3.1 Synthetic Data

A dataset of synthetic DCE-MRI sequences that simulate the behavior of a DCE-MRI acquisition was generated. The simulation was based on a two-dimensional 300x300 voxels phantom with separate regions representing the different tissues normally present in the anatomical region of the wrist. For each tissue a specific template defined the curve of the enhancement with respect to acquisition time, following [10]. The curves templates are shown in Fig. 1. Four sequences with 200 frames were generated, each for the following noise levels in $PSNR = \{21.82, 16.11, 12.89, 10.87\}$. During all the following experiments, the dictionary had $k = 50$ atoms and was learned with sparsity coefficient $\tau = 10$: the parameters were chosen in order to obtain a good trade-off between the sparsity of the matrix A and the PSNR performances. The average number of atoms used by the representations was 13.3.

Denoising performances. The first effect of sparse coding was to denoise the data. The PSNR over all the ECs is computed. The noisy ECs achieved a PSNR

⁴ <http://slipguru.disi.unige.it/research/PADDLE>

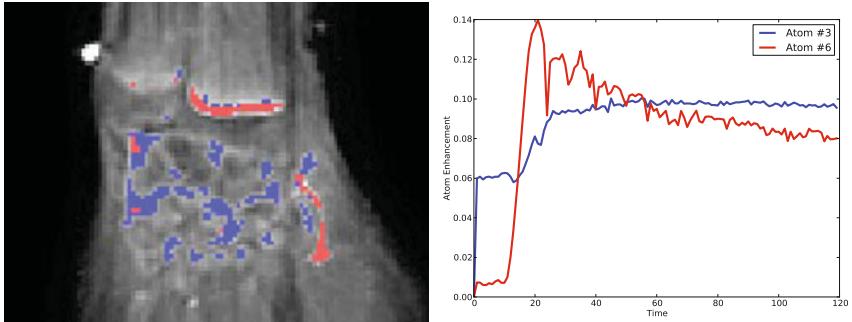


Fig. 2. Left: segmentations of two different type of tissues using matrix A of sparse codes on real data. Right: most used atoms related to the cluster; colors match in segmentation and atoms: vessels (red) and synovial volume (blue).

= 21.82 dB, while PSNR over denoised ECs is 31.46 dB. Moreover, the DL technique also achieved better performances compared to approaches based on polynomial fitting: the achieved PSNR using a third-degree polynomial curve fitting, as implemented in [1], is 27.28 dB.

Signal Recovery. As well as providing good denoising performance, the proposed method is also capable of recovering elementary signals that are good representatives of the different types of EC patterns. Figure 1 shows the most representatives atoms that are used in reconstructing the synthetic data, which correspond to the EC patterns of the phantom region. The good behaviour in the recovery of elementary signals is crucial to obtain suitable data representations that perform well when used for supervised and unsupervised learning tasks.

Clustering Results. The synthetic data represented in terms of the learned dictionary were clustered by k -means, with $k = 5$ clusters, one for each type of tissue: classification accuracy is reported in Table 1. Clustering the data with their sparse representations is nearly perfect, and outperforms the results obtained with both the raw values and the regression coefficients of a third-degree polynomial curve, as is done in [1].

3.2 Real Data

In order to test the effectiveness of the proposed method on real images, two different datasets were used, both acquired using an Intera Achieva 1.5T MR system (Philips Medical Systems) at the Istituto Pediatrico Giannina Gaslini (Genova, Italy). The first dataset consisted of 12 wrist DCE-MRI (Coronal 2D FFE; Resolution: 160x160; Slice thickness: 10mm; Pixel spacing: 1.06mm; 60 or 120 timepoints, depending on the acquisition) acquired from pediatric patients with active JIA. The second dataset consisted of 12 kidneys DCE-MRI

Table 1. Left: classification performances on synthetic data of k-means on original data matrix X , on sparse codes matrix A and on regression coefficients matrix, varying σ for Gaussian noise. Accuracy is defined as follows: $Acc. = 100 * \frac{(TP+TN)}{N}$ where TP is the number of True Positives, TN is the number of True Negatives and N is the total number of voxels. Right: comparison of denoising performances and computational time between our method and the two other approaches [1,7] using wrists dataset.

Acc.(%)	$\sigma = 5$	$\sigma = 10$	$\sigma = 15$	$\sigma = 20$	Wrists	PSNR(dB)	$\sigma(dB)$	Time(s)	$\sigma(s)$
X	77.9	77.6	77.1	77.1	DL	36.9	4.1	14.1	2.4
A	100	99.99	98.4	96.6	Agner	34.1	4.2	12.2	2.4
Regr.	71.3	66.2	59.1	53.95	Guo	18.75	6.1	305.3	75.82

(Coronal 3D FFE; Resolution 256x256x9; Slice thickness: 7mm; Pixel spacing: 1.33mm; 108 timepoints) acquired from pediatric patients with RD. DCE-MRI acquisitions of JIA and RD are proven to be effective in early diagnosis and disease monitoring for both pathologies [6,16]. In the experiments on the wrists dataset, the enhancing voxels were selected as those with $\bar{x}_i \geq 20$ and $\sigma_i \geq 3$ and belonging to a ROI automatically centered on the wrist based on heuristics regarding its orientation and size. The size of the dictionary k and the sparsity parameter τ were set to yield the sparsest representation achieving a reconstruction $PSNR \geq 25$ dB. The reported results have been obtained with $k = 20$ atoms and the sparsity around 40% of non-zero coefficients. In the experiments on the kidneys dataset, different settings were used. The automatic pruning of non-enhancing voxels was not reliable due to the large movements of the patients during the acquisitions: these strong variations may induce false negatives resulting in a wrong pruning of enhancing voxels. The voxels were selected using an approximate ROI that is manually placed, as shown in Figure 3, in order to reduce the computational time. All the reported clustering results has been obtained with $k = 10$ atoms and the sparsity around 40% of non-zero coefficients.

Wrists Dataset

Denoising performances. Table 1 shows that our method achieves better PSNR, than the other methods compared with [1,7]. The computational time for each patient, expressed as the total time elapsed to compute atoms and representations using our method and to compute the regression coefficients using [1,7] is reported: computation is done for all the ECs present in each patient.

Clustering results. Results of the clustering process are shown in Figure 2, expressed in terms of unsupervised classification, due to the lack of labeled real data. K-means algorithm was initialized with $k = 6$ clusters, which is the number of different tissue types typically present in the anatomical region of the wrist. The atoms found are good representatives of the different types of tissue normally found in anatomical region of the wrist, following Lavini et al. [10].

Kidneys Dataset

Clustering results. K-means clustering algorithm is applied on both sparse codes matrix A and original matrix X, in order to test the discriminative power of ECs

Table 2. Dice metric performances over Kidneys dataset, for Left(L.), Right(R.) and for both(L.+R.) kidneys. The segmentations were computed using sparse codes matrix (A), on table left, and original data matrix (X) as reported in Zöellner et al.[18], on table right.

Kidneys(A)	L.	R.	L.+R.	Kidneys(X)	L.	R.	L.+R.
Mean	0.883	0.783	0.830	Mean	0.827	0.746	0.784
Std. Dev.	0.059	0.086	0.089	Std. Dev.	0.085	0.113	0.106

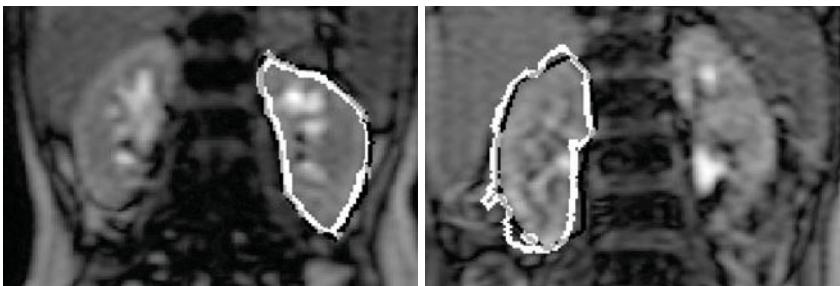


Fig. 3. Qualitative assessment of k-means clustering: superimposition between manual segmentation(black) and automatic segmentation(white) using matrix A of sparse codes on different patients. Overlapping signatures are signed in gray. Left: left kidney. Right: right kidney.

representation versus original ECs in the task of kidneys segmentation. As reported in Table 2, a comparison between the automatic segmentation to manual delineated (by a trained expert) regions is done in order to quantify the segmentation error: Dice overlap measure is applied [5]. The performances achieved by the proposed method are comparable with ones based only on the original ECs, as implemented in [18]. Figure 3 shows qualitative results of the clustering process: a superimposition between manual and automatic segmentation is done, showing the good behaviour of the proposed method in kidneys discrimination task.

4 Conclusions

This paper presented an automated method for processing and analyzing DCE-MRI sequences using a specific dictionary learning technique. The effectiveness and the efficiency of this novel approach is showed by presenting the results achieved in the specific tasks of signal recovery, denoising and unsupervised segmentation on both synthetic and real datasets of DCE-MRI acquisitions. The results are very encouraging, outperforming ones based on conventional techniques of curve analysis approach. Moreover, the presented method is extendable without efforts on different types of clinical context: in all these contexts, good performances were achieved. Future works will include the clinical validation of

the proposed method and its extension and adaptation to other clinical problems, such as prostate and breast cancer, in which particular attention will be paid to the problem of pharmacokinetic parameters extraction, as in [14].

References

1. Agner, S., et al.: Segmentation and classification of triple negative breast cancers using DCE-MRI. In: Proc. IEEE ISBI 2009, pp. 1227–1230 (2009)
2. Aharon, M., et al.: K-SVD: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on Signal Processing* 54(11) (2006)
3. Alonzi, R., Padhani, A.R., Allen, C.: Dynamic contrast enhanced MRI in prostate cancer. *Eur. J. Radiol.* 63(3), 335–350 (2007)
4. Bishop, C.: *Pattern Recognition and Machine Learning*. Springer, Heidelberg (2006)
5. Crum, W., et al.: Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE TMI* 25(11), 1451–1461 (2006)
6. Damasio, M.B., Malattia, C., Martini, A., Tomà, P.: Synovial and inflammatory diseases in childhood: role of new imaging modalities in the assessment of patients with juvenile idiopathic arthritis. *Pediatric Radiology* 40(6), 985–998 (2010)
7. Guo, J., Reddick, W.: DCE-MRI pixel-by-pixel quantitative curve pattern analysis and its application to osteosarcoma. *Journal of MR* 30(1), 177–184 (2009)
8. Harris, N., Gauden, V., Fraser, P., Williams, S., Parker, G.: MRI measurement of blood-brain barrier permeability following spontaneous reperfusion in the starch microsphere model of ischemia. *Magnetic Resonance Imaging* 20(3), 221–230 (2002)
9. Kubassova, O., Boesen, M., Boyle, R.D., Cimmino, M.A., Jensen, K.E., Bliddal, H., Radjenovic, A.: Fast and robust analysis of dynamic contrast enhanced MRI datasets. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) *MICCAI 2007, Part II. LNCS*, vol. 4792, pp. 261–269. Springer, Heidelberg (2007)
10. Lavini, C., et al.: Pixel-by-pixel analysis of DCE MRI curve patterns and an illustration of its application to the imaging of the musculoskeletal system. *Magnetic Resonance Imaging* 25(5), 604–612 (2007)
11. Lee, H., Battle, A., Raina, R., Ng, A.: Efficient sparse coding algorithms. In: *Advances in Neural Information Processing Systems, NIPS 2006*, vol. 19 (2006)
12. Mairal, 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. Olshausen, B., Field, D.: Sparse coding with an overcomplete basis set: A strategy employed by V1? *Vision Research* 37(23), 3311–3325 (1997)
14. Schmid, V.J., et al.: Quantitative analysis of dynamic contrast-enhanced MR images based on bayesian p-splines. *IEEE TMI* 28(6), 789–798 (2009)
15. Stagljanò, A., Chiusano, G., Bassi, C., Santoro, M.: Learning adaptive and sparse representations of medical images. In: Menze, B., Langs, G., Tu, Z., Criminisi, A. (eds.) *MICCAI 2010. LNCS*, vol. 6533, pp. 130–140. Springer, Heidelberg (2011)
16. Vivier, P., Blondiaux, E., Dolores, M., Marouteau-Pasquier, N., Brasseur, M., Petitjean, C., Dacher, J.: Functional mr urography in children. *J. Radiol.* (2009)
17. Yang, J., Yu, K., Gong, Y., Huang, T.: Linear spatial pyramid matching using sparse coding for image classification. In: *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2009*, pp. 1794–1801 (June 2009)
18. Zöllner, F.G., et al.: Assessment of 3D DCE-MRI of the kidneys using non-rigid image registration and segmentation of voxel time courses. *Computerized Medical Imaging and Graphics* 33(3), 171–181 (2009)

Learning Optical Flow Propagation Strategies Using Random Forests for Fast Segmentation in Dynamic 2D & 3D Echocardiography

Michael Verhoek, Mohammad Yaqub, John McManigle, and J. Alison Noble

Institute of Biomedical Engineering, University of Oxford

Abstract. Fast segmentation of the left ventricular (LV) myocardium in 3D+time echocardiographic sequences can provide quantitative data of heart function that can aid in clinical diagnosis and disease assessment. We present an algorithm for automatic segmentation of the LV myocardium in 2D and 3D sequences which employs learning optical flow (OF) strategies. OF motion estimation is used to propagate single-frame segmentation results of the Random Forest classifier from one frame to the next. The best strategy for propagating between frames is learned on a per-frame basis. We demonstrate that our algorithm is fast and accurate. We also show that OF propagation increases the performance of the method with respect to the static baseline procedure, and that learning the best OF propagation strategy performs better than single-strategy OF propagation.

1 Introduction

Quantitative determination of wall motion and blood pool volume of the cardiac left ventricle (LV) during the heart cycle can aid assessment of heart abnormalities and heart disease. Fast and accurate automatic segmentation of the LV myocardium can provide such quantitative data from 3D echocardiographic sequences. However, segmentation in 3D echocardiography is challenging due to speckle, missing boundaries and different tissues with similar appearance. Furthermore, dense 3D+time sequences provide a large amount of data to be processed; given that acquisition is real-time, it is highly desirable that analysis methods are too.

Segmentation of ultrasound images has previously been performed by traditional methods like level sets and active contours [1]. Accordingly, such methods have been applied to LV endocardial segmentation in 3D echocardiography [2,3,4]. The epicardial boundary is more challenging to find than the endocardial boundary due to the presence of nearby tissues, hence research on myocardial segmentation is more sparse [5,6]. These segmentation methods have in common that they are not near-real-time.

This limitation is overcome in our work by using a discriminative classifier for the segmentation task, where each pixel/voxel is classified as either myocardium or non-myocardium. Discriminative classifiers do not model the data, but are

trained to learn the classification task directly from the input data; the lack of (restrictive) assumptions about the data also leads to a simpler approach than competing algorithms. The specific classifier we use is the Random Forest (RF) classifier [7,8], which uses fewer assumptions than other discriminative classifiers like SVMs or neural networks. The simplicity of the RF classifier means it can run very efficiently. Its use in segmentation in medical imaging has been quite recent, and includes neuron segmentation in electron-microscopy [9], brain tissue segmentation in MRI [10], myocardial segmentation in static 3D echocardiography [11] and MS lesion segmentation in brain MRI [12].

Our work is most closely related to [11] in which the authors apply an RF for fast segmentation in static 3D images. The novel contribution of this paper is the use of optical flow (OF) motion cues on static RF segmentations and learning the best-performing OF propagation strategy for each frame, enabling the application of RFs for the first time to dynamic 3D+time echocardiography.

2 Method

2.1 Data Sets

Twenty five 3D+t apical view echocardiograms were obtained from the JR Hospital in Oxford. The subjects were all healthy with each study recorded on a Philips iE33 ultrasound system by a cardiologist. Image spatial resolution was on average $0.88 \times 0.89 \times 0.81 \text{ mm}^3$ and sequence temporal resolution was on average 59 ms; each sequence consisted of between 11 and 20 time frames, starting at the systolic phase. The first 11 frames of each sequence were used in the experiments. The scans showed significant variability in brightness so histogram equalization was performed as a preprocessing step.

Experiments were performed on either (1) the 3D+t dataset, or (2) a 2D+t dataset, both from 25 scans each. In both cases, manual segmentations were performed by two experts: in each of the 25 3D sequences, one of the experts delineated the LV myocardium in every fifth short-axis slice of each volume. The expert delineated an epicardial and an endocardial contour in each fifth short-axis slice, where the orthogonal long-axis views were also visible to review the segmentation in those views. Papillary muscles and valves were considered to belong to the blood pool, and in places where the myocardium was less visible due to shadowing, attenuation or a parallel beam, expert knowledge of the shape of the heart wall was used to delineate the boundary. In the 2D+t case, for each of the 25 sequences one short-axis slice per frame was considered. The short-axis slice was selected, located at one third of the distance between base and apex, from the base. See fig. 1 for an example of the dataset and manual segmentation.

2.2 Random Forests

Random Forests [7,8] is an ensemble of binary decision trees [13] and follows from earlier ensemble methods like bagging and earlier tree methods like probabilistic boosting trees (PBTs), leading to a smaller generalization error and increased

speed [14]. A Random Forest consists of a collection of binary decision trees. The trees are constructed during the training phase, in which every non-leaf node in the tree is assigned a binary test, which directs a training data sample to one of the two child nodes. All training data passes from the root through the tree and ends up at a leaf node, where the distribution of training data reaching the leaf node is kept. During the testing phase, a new data sample is passed down the trees to end up at a leaf node, where it is assigned the distribution of that leaf node; the forest output is the average of all distributions at the leaf nodes reached in each tree.

Random Forests are random in two ways. First, of the training data with size N_{tr} , a random subset $\varepsilon N_{\text{tr}}$ is available for each tree. Second, for each tree node, the algorithm chooses the best binary test from a set of randomly generated candidate tests N_{tests} .

We follow the setup of the Random Forest algorithm described in [11]; here we briefly mention some important points. The training set consists of all pixels¹ in the 2D or 3D images. A forest consist of N_{tree} trees. The maximum tree depth is D_{\max} , and N_{θ} thresholds are generated and tested for each candidate test. The best test and threshold is determined according to the maximum information gain measure. Two test types are used to keep the algorithm simple and fast: appearance and location. The appearance test sums the intensity values in a box centered at the pixel under consideration. The box sizes are randomly chosen for each candidate test, but with a maximum size R_{\max} ($R_{\max}^{x,y,z}$ are equal unless stated otherwise). The position test looks at the global absolute location of pixels. A dimension is picked at random and the current pixel's location in that dimension is compared to a threshold. By using one test type as a child of another, these test types can be combined to capture more complex local features. The output of running a trained Random Forest on a testing image is a soft segmentation (i.e. consisting of myocardium probabilities) RF , that can be thresholded to obtain a binary segmentation.

2.3 Optical Flow

In cases where the appearance and position information do not yield satisfactory results, for instance due to missing walls, motion cues can be employed. Speckle tracking can be used, but requires the radiofrequency signal, which conflicts with the requirement of a fast processing solution. We have therefore chosen to work with the B-mode and use optical flow (OF). Using OF, soft segmentation results from one frame can be propagated to another frame. In this work we measure the displacement based on a combined local global (CLG) method, a 3D version of Bruhn's 2D optical flow algorithm [15]. The calculated OF between consecutive frame pairs in a 2D and 3D sequence is then used after myocardial segmentation using the RF algorithm, to propagate the segmentation result from one frame to another. In our work, several OF propagation strategies are compared against each other, but a novelty of our work is that we learn the best propagation strategies.

¹ Note that we use "pixel" for both 2D and 3D picture elements.

If a sequence S consists of images $I_{S,fr}$, where $fr \in (1\dots11)$ is the frame number, then optical flow motion fields $OF(I_{S,fr}, I_{S,fr+1})$ are calculated from the intensity values in all consecutive image pairs. These motion fields are used to warp or propagate $RF_{S,fr}$ to $RF_{S,fr}^{OF(fr \rightarrow fr+1)} \equiv \hat{RF}_{fr+1}^{fr}$, where the superscript denotes the origin frame and the subscript the target frame. Reverse ($\hat{RF}_{fr}^{fr+1} = RF_{S,fr}^{OF(fr+1 \rightarrow fr)}$) and cumulative ($\hat{RF}_{fr+2}^{fr} = RF_{S,fr+2}^{OF(fr \rightarrow fr+2)}$) propagations are also possible by using negative and summed motion fields respectively. We want to test all possible origin frame–target frame propagations. To this end, a system of propagation strategies is set up, based on the origin frame and propagating to all other frames. For example, strategy $str = o1$ (“origin frame 1”) is illustrated in fig. 2 and consists of a sequence where $RF_{S,fr=1}$ is propagated to the other frames: $\hat{RF}_S^{str=o1} = (RF_{S,1}, \hat{RF}_{S,2}^1, \dots, \hat{RF}_{S,11}^1)$, where the vector runs over the 11 frames of this new segmentation sequence. Note that if the origin and target frame are equal (frame 1 in the example), the original, non-propagated RF result is used. This ensures that the non-propagated results are compared against the propagated results. In this way, propagated segmented sequences \hat{RF}_S^{str} are obtained for all strategies $str \in \{o1, o2, \dots, o11\}$. The strategies are illustrated in fig. 2. The next step is to increase the number of strategies by combining propagations in the 11 basic strategies, for example $str = o2 + o6$ means the average of propagations coming from frame 2 and from frame 6: $\hat{RF}_S^{str=o2+o6} = (\frac{1}{2}(\hat{RF}_{S,1}^2 + \hat{RF}_{S,1}^6), \frac{1}{2}(RF_{S,2} + \hat{RF}_{S,2}^6), \dots, \frac{1}{2}(\hat{RF}_{S,11}^2 + \hat{RF}_{S,11}^6))$. There are 55 combined strategies possible, yielding a total number of 66 strategies.

This large number of OF propagation strategies are attempted, with the goal to learn the best performing strategy for each frame. Suppose there is a measure $y(str, fr)$ which denotes the performance of strategy str for frame fr , for all sequences S together, i.e. $y(1, fr) > y(2, fr)$ means strategy 1 performs better than strategy 2 for frame fr , over all sequences. Then we find, for $fr = 1$, $str_1^* = \arg \max_{str} y(str, 1)$. This is repeated for all frames to obtain the best strategy sequence $str^* = (str_1^*, str_2^*, \dots, str_{11}^*)$. This best strategy is applied to an RF segmented sequence to obtain the best segmentation: $\hat{RF}_S^{str^*} = (\hat{RF}_{S,1}^{str_1^*}, \dots, \hat{RF}_{S,11}^{str_{11}^*})$.

To segment a new, previously unseen, 3D sequence, its individual frames would be passed through each frame’s trained RF and the result would be OF propagated according to the target frame’s learned best OF strategy.

2.4 Experiments

Baseline experiments. The performance of the (static) 2D and 3D RF segmentation algorithms were tested on all 25 2D and 3D sequences. This was done on a per-frame basis, i.e. an RF was trained and tested for each of the 11 temporal frames separately. Leave-one-out cross-validation (LOOCV) was performed, i.e. the RF was trained on 24 images and tested on the remaining image, repeated for all 25 cases, and then repeated for all 11 frames, leading to soft segmentation

(probability) results $RF_{S,fr}$ ($S = 1 \dots 25, fr = 1 \dots 11$). To get a hard segmentation, the probability image is thresholded. A range of probability thresholds were used and the resulting segmentation compared with the manual segmentations $M_{S,fr}$ to obtain ROC curves, as a function of the probability threshold θ , $ROC_{S,fr}(\theta)$. The results are accumulated along S for fixed fr to obtain $ROC_{fr}(\theta)$. The area under each ROC curve, A_{fr} , is also computed; both are measures of the performance of the algorithm for that frame. For overall performance measures, the results are accumulated along fr to obtain $ROC(\theta)$ and A .

2D+t volume segmentation. As a comparison, RF segmentation was then applied to the 25 2D+t sequences by treating the time dimension as the third dimension. No OF propagations were used. In this case the maximum box size R_{\max} was changed in the time dimension to $R_{\max}^t = 2$. $R_{\max}^{x,y}$ were kept at 19. Again, LOOCV was performed and ROC curves obtained as described above.

OF propagation. We used OF to perform dynamic segmentation in 2D and 3D. In each case, after obtaining the RF results for each image frame from the static baseline experiments, the probabilistic segmentation results were propagated to other frames according to the strategies described in Sec. 2.3. After obtaining soft segmentation results for each strategy $\hat{RF}_{S,fr}^{str}$, these were compared with $M_{S,fr}$ to obtain ROC curves for each strategy, which in turn were accumulated along S to obtain $ROC_{fr}^{str}(\theta)$ and A_{fr}^{str} , measures of the performance per frame, per strategy.

The best-performing OF propagation strategy was learned, across all images: A_{fr}^{str} was used as the performance measure $y(str, fr)$ described above, to obtain a “best-of” strategy sequence. LOOCV was employed to test the performance: 24 sequences were taken together to obtain $ROC_{fr}^{str}(\theta)$ and A_{fr}^{str} , yielding a best-of sequence to be applied to the remaining segmentation sequence RF_S ; this was repeated 25 times to obtain 25 $\hat{RF}_S^{str^*}$; accumulation along S and fr yields ROC^{str^*} .

The values for the global parameters of the algorithm were set on a subset of 5 out of 25 images by varying one while keeping the others fixed and comparing the results. The parameters in question were the maximum tree depth $D_{\max} = 16$, the number of candidate tests at each node $N_{\text{tests}} = 30$ and the maximum rectangle dimension $R_{\max} = 32$ for the 3D case and $R_{\max} = 19$ for the 2D case. Other variables kept fixed throughout were $N_\theta = 8$, $\varepsilon = 0.25$. The number of trees was kept at $N_{\text{tree}} = 3$, as previous work has shown that a larger number of trees does not improve results significantly, while increasing run-time. The code used for our experiments is based on the algorithm in [16], adapted for use on 3D images.

3 Results and Discussion

For the 2D and 3D baseline experiments, accumulative ROC curves (summed for all images and frames) are presented fig. 3a and 3b. Fig. 3a shows an accumulative ROC curve of the stacked 2D+t experiment. A_{fr}^{str} for the OF propagation experiment is shown in fig. 4a for the 11 basic strategies and in fig. 4b for the four best performing combination strategies (as well as the 2 worst performing ones; the remaining 49 are in between). These lines show the A_{fr}^{str} for all 25

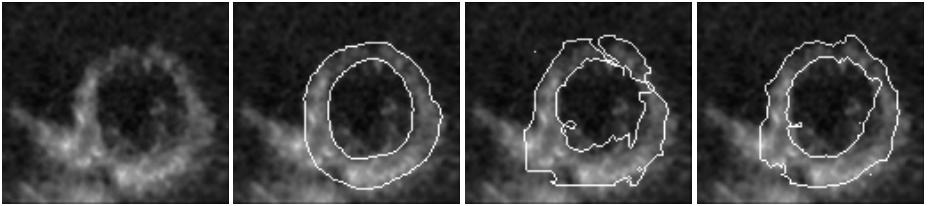


Fig. 1. Sample short-axis slices of (left to right) the data set, manual segmentation, thresholded static RF segmentation result, thresholded best OF-propagated segmentation result (“neighbor”). Frame 8, z=90.

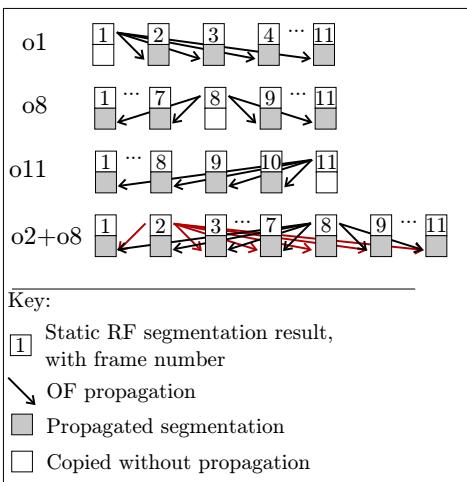


Fig. 2. Diagrams of example optical flow propagation strategies: 3 basic and 1 combination. Note: for “o₂+o₈”, both incoming propagations are averaged).

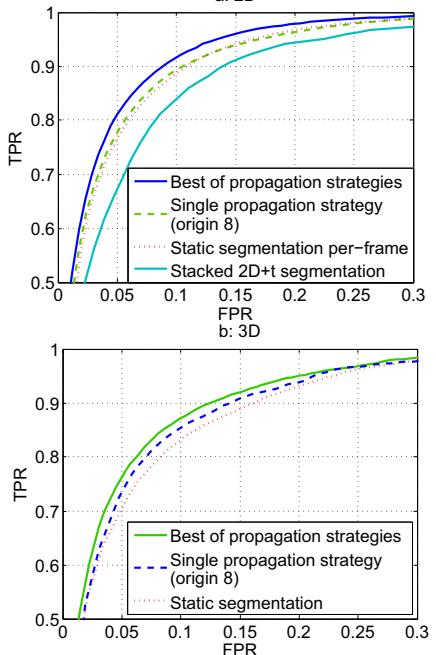


Fig. 3. ROC curves for 2D (a) and 3D (b)

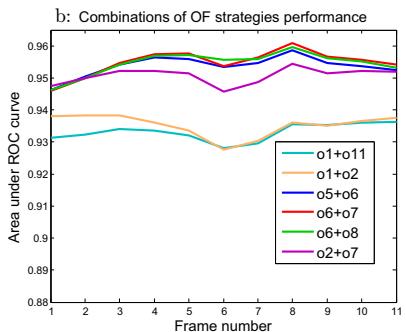
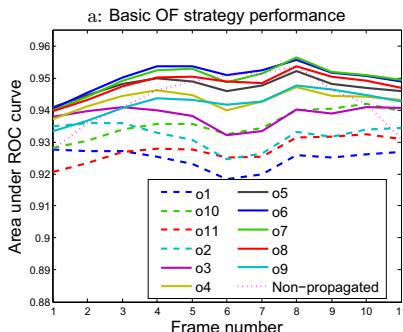


Fig. 4 Area under the ROC curves for all sequences, for (a) each of the basic OF propagation strategies and the non-propagated RF segmentation result and (b) the four best and two worst of the combination strategies, as a function of the frame number. Best results: frame 1: o₂+o₇; frame 2: o₅+o₆; frame 6: o₆+o₈; other frames: o₆+o₇.

sequences accumulated, rather than one of the LOOCV cases, but each of the LOOCV graphs is very similar to the graphs shown here. As an illustration, the performance of a single propagation strategy is shown as an accumulated ROC curve in fig. 3a and 3b (both o8 strategy). Finally, the performance of the “best-of” segmentation result sequence is shown as an accumulated ROC curve (all LOOCV cases accumulated), plotted in fig. 3a for 2D and fig. 3b for 3D. Fig. 4 also makes clear what the learned best-of strategy is, if we were to apply the learned strategy to a new sequence: the best segmentation of for example the last five frames is o6+o7, while the best segmentation of frame 6 is o6+o8. In general, the newly proposed Learned Propagation Strategy (LPS) approach for myocardial segmentation yields a TPR of 92% at FPR=10% in 2D, and in 3D a TPR=87% at FPR=10%, compared to TPR=83% for the static baseline experiment and TPR=85% for the example single strategy. The 2D result performs better than the 3D result, due to the larger variability in shapes and the presence of more tissues with similar appearance in 3D.

As can be seen in fig. 3a, a 2D per-frame approach performs better than a stacked 2D+t approach. This shows that the RF does not pick up on correlations in the time-dimension, possibly due to larger displacements in the time dimension. This justifies our use of the per-frame approach also in the 3D case.

Fig. 3 shows that in both 2D and 3D, a single propagation strategy performs better than none at all. Furthermore, a “best-of” propagation strategy performs better than a single propagation strategy. Fig. 4 shows that, of the basic OF propagation types, o6, o7 and o8 perform well. These frames are the end-systolic frames in almost all sequences; in this phase, heart walls are most likely to be visible, as the muscle is at its highest density. Combination strategies perform better in general than basic strategies because two-frame interpolation gives better results than results coming from just one frame. Of these, combinations of the above mentioned end-systolic frames perform best, such as o6+o7.

On a current 2.4 GHz computer, segmenting one 3D frame takes about 2s. With our implementation, this is not fast enough to run in real-time. Furthermore, it should be noted that calculating an optical flow displacement image takes about 2 minutes per frame, destroying our fast results in theory. However, RFs have been shown to work very fast on a GPU implementation [17], enabling real-time RF application; also, optical flow is not intrinsically a slow method, OF implementations on a GPU have shown a 150-fold increase [18].

4 Conclusion

We have developed a novel RF-based algorithm for myocardial segmentation of 2D+t and 3D+t echocardiographic images. Motion cues were used in the form of optical flow propagated segmentation results. The use of a single optical flow propagation strategy boosted the performance slightly in both 2D and 3D. Moreover, learning the best combined propagation strategy boosted the performance further. We found that the segmentation is accurate (88% TPR at 10% FPR), and the RF segmentation is fast. Implementation of our algorithm on a GPU chip is potentially (near-)real-time.

Future work includes the use of images features other than intensity as an input for the RFs, a larger dataset, the combination of more than two strategies and smarter combination than averaging, and finally, implementation of our algorithm on a GPU to enable real-time solution.

References

- Noble, J., Boukerroui, D.: Ultrasound image segmentation: a survey. *IEEE Trans. Med. Imaging* 25(8), 987–1010 (2006)
- Corsi, C., Saracino, G., et al.: Left ventricular volume estimation for real-time three-dimensional echocardiography. *IEEE TMI* 21(9), 1202–1208 (2002)
- Angelini, E.D., Homma, S., Pearson, G., Holmes, J.W., Laine, A.F.: Segmentation of real-time three-dimensional ultrasound for quantification of ventricular function. *Ultrasound Med. Biol.* 31(9), 1143–1158 (2005)
- Leung, K.E., Bosch, J.G.: Automated border detection in three-dimensional echocardiography. *Eur. J. Echocardiogr.* 11(2), 97–108 (2010)
- Zhu, Y., Papademetris, X., Sinusas, A.J., Duncan, J.S.: A coupled deformable model for tracking myocardial borders from real-time echocardiography using an incompressibility constraint. *Med. Image Anal.* 14(3), 429–448 (2010)
- Myronenko, A., Song, X., Sahn, D.J.: LV motion tracking from 3D echocardiography using textural and structural information. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) *MICCAI 2007*, Part II. LNCS, vol. 4792, pp. 428–435. Springer, Heidelberg (2007)
- Breiman, L.: Random forests. *Machine Learning* 45(1), 5–32 (2001)
- Amit, Y., Geman, D.: Shape quantization and recognition with randomized trees. *Neural Comput.* 9(7), 1545–1588 (1997)
- Andres, B., Köthe, U., Helmstaedter, M., Denk, W., Hamprecht, F.A.: Segmentation of SBFSEM volume data of neural tissue by hierarchical classification. In: Rigoll, G. (ed.) *DAGM 2008*. LNCS, vol. 5096, pp. 142–152. Springer, Heidelberg (2008)
- Yi, Z., Criminisi, A., Shotton, J., Blake, A.: Discriminative, semantic segmentation of brain tissue in MR images. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) *MICCAI 2009*. LNCS, vol. 5762, pp. 558–565. Springer, Heidelberg (2009)
- Lempitsky, V., Verhoek, M., Noble, J.A., Blake, A.: Random forest classification for automatic delineation of myocardium in real-time 3D echocardiography. In: Ayache, N., Delingette, H., Sermesant, M. (eds.) *FIMH 2009*. LNCS, vol. 5528, pp. 447–456. Springer, Heidelberg (2009)
- Geremia, E., Menze, B.H., Clatz, O., Konukoglu, E., Criminisi, A., Ayache, N.: Spatial decision forests for MS lesion segmentation in multi-channel MR images. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) *MICCAI 2010*. LNCS, vol. 6361, pp. 111–118. Springer, Heidelberg (2010)
- Quinlan, J.R.: Induction of decision trees. *Machine Learning* 1, 81–106 (1986)
- Caruana, R., Niculescu-Mizil, A.: An empirical comparison of supervised learning algorithms. In: ICML 2006, pp. 161–168. ACM, New York (2006)
- Bruhn, A., Weickert, J., Schnörr, C.: Lucas/Kanade meets Horn/Schunck: combining local and global optic flow methods. *Int. J. Comp. Vision* 61(3), 1–21 (2005)
- Shotton, J., Johnson, M., Cipolla, R.: Semantic texton forests for image categorization and segmentation. In: CVPR 2008, pp. 1–8 (2008)
- Sharp, T.: Implementing decision trees and forests on a GPU. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) *ECCV 2008*, Part IV. LNCS, vol. 5305, pp. 595–608. Springer, Heidelberg (2008)
- Pauwels, K., Hulle, M.V.: Realtime phase-based optical flow on the GPU. In: CVPRW 2008, pp. 1–8 (2008)

A Non-rigid Registration Framework That Accommodates Pathology Detection

Chao Lu¹ and James S. Duncan^{1,2}

¹ Department of Electrical Engineering, School of Engineering & Applied Science

² Department of Diagnostic Radiology, School of Medicine

Yale University, New Haven, CT, USA

chao.lu@yale.edu

Abstract. Image-guided external beam radiation therapy (EBRT) for the treatment of cancer enables accurate placement of radiation dose to the cancerous region. However, the deformation of soft tissue during the course of treatment, such as in cervical cancer, presents significant challenges. Furthermore, the presence of pathologies such as tumors may violate registration constraints and cause registration errors. In this paper, we present a novel MAP framework that performs nonrigid registration and pathology detection simultaneously. The matching problem here is defined as a mixture of two different distributions which describe statistically image gray-level variations for two pixel classes (i.e. tumor class and normal tissue class). The determinant of the transformation's Jacobian is also constrained, which guarantees the transformation to be smooth and simulates the tumor regression process. We perform the experiments on 30 patient MR data to validate our approach. Quantitative analysis of experimental results illustrate the promising performance of this method in comparison to previous techniques.

1 Introduction

Cervical cancer is a common type of cancer in women. The traditional treatment for carcinoma of the cervix has been by surgery or external beam radiation therapy(EBRT), however, when the disease is on an advanced stage, surgery is not possible, and EBRT is the primary modality of treatment. Radiotherapy is feasible, effective, and is used to treat over 80% of patients [1].

While CT images are commonly used for radiotherapy, for cervical cancer however, T2-weighted MR imaging is increasingly added, because MRI provides superior visualization and performance for measuring cervical carcinoma size and uterine extension[2]. In addition, advanced MR imaging with modalities such as diffusion, perfusion and spectroscopic imaging has the potential to better localize and understand the disease and its response to treatment. Therefore, magnetic resonance guided radiation therapy (MRgRT) systems with integrated MR imaging in the treatment room are now being developed as an advanced system in radiation therapy[3,4,5].

During radiation therapy treatment, a patient is irradiated multiple times across several weeks. The main factor that affects the success of treatment is

dose targeting, i.e. to deliver as much dose as possible to the clinical target volume (CTV), while trying to deliver as little dose as possible to surrounding normal organs at risk[6]. Furthermore, the prescribe treatment plans need to be adapted to changes in the anatomy due to the unpredictable inter- and intra-fractional organ motions, which calls for an accurate mapping and anatomical correspondence between treatment and planning day. However, the existence of pathologies such as tumors may violate registration constraints and cause registration errors, because the abnormalities often invalidate gray-level dependency assumptions that are usually made in the intensity-based registration. In addition, the registration problem is challenging due to the missing correspondences caused by tumor regression. Learning some knowledge about abnormalities and incorporating them into the framework can improve the registration.

The automatic mass detection problem mainly focuses on the application on mammography[7]. Meanwhile, the registration problem is usually addressed by minimizing the energy function defined using intensity similarity metric, such as in [8,9]. In [4], the authors present a multifeature mutual information based method to perform registration of cervical MRI. However, as above discussed, tumorous tissue may disappear in time due to successful treatment, which distorts the intensity assumptions and raises questions regarding the applicability of these approaches.

In this paper, we present a novel intensity-based registration framework to nonrigidly align intrapatient MR images of the cervix, for the motivations mentioned above. We define the matching problem as a mixture of two different distributions which describe statistically image gray-level variations for two pixel classes (i.e. tumor class and normal tissue class). These mixture distributions are weighted by the tumor detection map which assigns to each voxel its probability of abnormality. We also constraint the determinant of the transformation's Jacobian, which guarantees the transformation to be smooth and simulates the tumor regression process. Using the proposed method, we are able to solve the challenging problems induced by the presence and regression of tumor, which makes the new technique more suitable for application in image guided radiotherapy and computer aided diagnosis.

2 Method

2.1 Bayesian Formulation

Let f and g be the planning day and treatment day 3D images respectively. Follow a Bayesian line of thinking, we estimate the deformation field between two image data T , and the tumor map M in treatment day which associates to each pixel of the image g its probability of belonging to a tumor.

$$\begin{aligned} \widehat{T}, \widehat{M} &= \arg \max_{T, M} [p(T, M | f, g)] \\ &= \arg \max_{T, M} [\ln p(f, g | T, M) + \ln p(T | M) + \ln p(M)] \end{aligned} \quad (1)$$

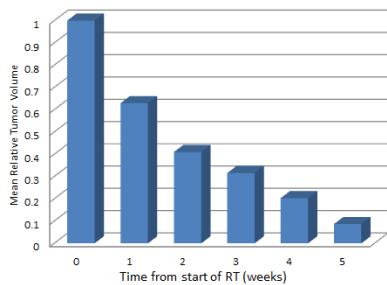


Fig. 1. Tumor regression: plot of mean relative tumor volume

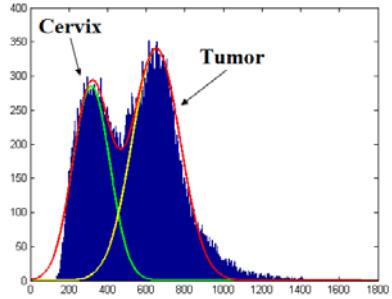


Fig. 2. The histograms of normal cervix and tumor

2.2 Intensity Matching and Tumor Probability Map

In order to define the likelihood $p(f, g|T, M)$, we assume conditional independence over the voxel locations (x), as discussed in [10].

$$p(f, g|T, M) = \prod_{(x)} p(T(f(x)), g(x)|M(x)) \quad (2)$$

Different from the previous work in [8,9,11,4] which only use a single similarity metric, here we model the probability of the pair $p(T(f(x)), g(x)|M(x))$ to be dependent on the class of the pixel (x). Each class is characterized by a probability distribution, denoted by p_N for the normal tissue and p_T for the tumor. Let $M(x)$ be the tumor map which associates to (x) its probability to belong to a tumor. Thus, the probability distribution $p(T(f(x)), g(x)|M(x))$ can be defined as a mixture of the two class distribution:

$$p(T(f(x)), g(x)|M(x)) = (1 - M(x))p_N(g(x), T(f(x))) + M(x)p_T(g(x), T(f(x))) \quad (3)$$

Normal Tissue Class. Across the treatment procedure, the tumor experiences a regression process if the treatment is successful [2]. The tumor regression is shown in Fig.1, which presents a plot of mean relative tumor volume as measured with weekly MRI during EBRT in 5 patients with cancer of the uterine cervix. When the tumor shrinks, some part of the tumor turns into scar and returns to the intensity level around the normal tissue. Thereafter, the normal tissue in treatment day MR has two origins: normal tissue in planning day MR, and tumor in planning day but returns to normal due to the tumor regression. We choose two different probability for these two types. The histograms of normal cervical tissue and tumor are shown in Fig.2. From clinical research[12] and Fig.2, we can see that the intensities of tumor are generally much higher than those of normal cervical tissue.

Therefore, for the sake of simplicity, we characterize the normal tissue from the second type (used to be tumor but returned to normal later on) as areas with

much lower intensity in treatment day MR[12]. We assume a voxel labeled as this part in day d MR can match any voxel in day 0 tumor with equal probability and use a uniform distribution. The remaining voxels are labeled as type 1 normal tissue (always normal since the planning day), which are modeled assuming a discrete Gaussian distribution across the corresponding voxel locations.

$$p_N(g(x), T(f(x))) = \begin{cases} 1/c, & T(f(x)) - g(x) > \Delta \\ \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{|T(f(x)) - g(x)|^2}{2\sigma^2}\right), & \text{otherwise} \end{cases} \quad (4)$$

where c is the number of voxels in the day 0 tumor, and Δ is the predetermined threshold used to differentiate the intensity of normal and tumor tissue.

Tumor Class. The definition of the tumor distribution is a difficult task. Across the radiotherapy the tumor regression process, the tumor in treatment day d is the remaining tumor after the radiation treatment. We assume the voxel in the residual tumor can match any voxel in the initial tumor (day 0) with equal probability and use a uniform distribution.

$$p_T(g(x), T(f(x))) = \frac{1}{c}, \quad (5)$$

where c is the number of voxels in the day 0 tumor.

Transformation Smoothness. We want the transformation to be non-singular hence we would like to have a large penalty on negative Jacobian for normal tissues. Meanwhile, we simulate the tumor regression process by constraining the determinant of the transformation's Jacobian at $M(x)$ between 0 and 1.

$$p(T(x)|M(x)) = (1 - M(x))p_N(T(x)) + M(x)p_T(T(x)) \quad (6)$$

where p_N and p_T can be calculated using the following continuous functions: ($\epsilon = 0.1$)

$$p_N(T_{0d}(\mathbf{x})) = \frac{1}{2} \left[1 + \frac{2}{\pi} \arctan\left(\frac{|J(T_{0d}(\mathbf{x}))|}{\epsilon}\right) \right] \quad (7)$$

$$p_T(T_{0d}(\mathbf{x})) = \frac{1}{2} \left[1 + \frac{2}{\pi} \arctan\left(\frac{|J(T_{0d}(\mathbf{x}))|}{\epsilon}\right) \right] - \frac{1}{2} \left[1 + \frac{2}{\pi} \arctan\left(\frac{|J(T_{0d}(\mathbf{x}))| - 1}{\epsilon}\right) \right] \quad (8)$$

These constraints (p_N and p_T) penalize negative Jacobians. p_T guarantees the determinant of the transformation's Jacobian less than 1, which simulates the tumor regression process, thus reduce the probability of folding in the registration maps.

Tumor Map Prior. We assume that the tumor map arises from a Gibbs distribution. Many specific terms can be defined to describe the spatial configurations of different type of lesion. The previous detection approaches have a tendency

to *over-detect* (to find more regions than the real ones)[7], hence in this paper, we use an energy restricting the total amount of abnormal pixels in the image.

$$p(M) = \frac{1}{Z} e^{-U(M)} = \frac{1}{Z} e^{-\sum_x M(x)} \quad (9)$$

Energy Function. Combining the above equations, we introduce the energy function:

$$\begin{aligned} E(T, M) &= -\ln p(T, M|f, g) \\ &= -\int_x \ln p(g(x), T(f(x))|M(x))dx - \int_x \ln p(T(x)|M(x))dx + \sum_x M(x) \end{aligned} \quad (10)$$

where $p(g(x), T(f(x))|M(x))$, $p(T(x)|M(x))$ can be replaced by Eq.(3) and Eq.(6) accordingly. The transformation T is represented by B-Spline Free Form Deformation[8]. When minimized, the transformation T as well as the tumor probability map M are estimated simultaneously.

3 Results

We tested our proposed method on 24 sets of MR data acquired from four different patients undergoing EBRT for cervical cancer at Princess Margaret Hospital, Toronto, Canada. Each of the patient had six weekly 3D MR images. A GE EXCITE 1.5-T magnet with a torso coil was used in all cases. T2-weighted, fast spin echo images (voxel size $0.36mm \times 0.36mm \times 5mm$, and image dimension $512 \times 512 \times 38$) were acquired. The Δ in Eq.(4) was chosen to be 150. The clinician performed the bias field correction so that the intensities could be directly compared, and the MR images were resliced to be isotropic.

3.1 Registration Results

The registration performance of the proposed algorithm was evaluated. For comparison, an ICM-based nonrigid registration [6], an intensity-based FFD nonrigid registration [8] and a rigid registration were performed on the same sets of real patient data.

The Dice coefficient between the ground truth in day d and the transformed organs from day 0 were used as metrics to assess the quality of the registration (Table 1). We also tracked the registration error percentage between the ground truth in day d and the transformed organs from day 0 for bladder and uterus, as shown in Table 2. The registration error is represented as percentage of false positives (PFP), which calculates the percentage of a non-match being declared to be a match. Let Ω_A and Ω_B be two regions enclosed by the surfaces A and B respectively, Dice coefficient and PFP are defined as follows:

$$Dice = \frac{2Volume(\Omega_B) \cap Volume(\Omega_A)}{Volume(\Omega_B) + Volume(\Omega_A)}; PFP = \frac{Volume(\Omega_B) - Volume(\Omega_A \cap \Omega_B)}{Volume(\Omega_A)}.$$

Table 1. Evaluation of Registration: Dice

	Bladder	Uterus
Rigid Registration	0.36 ± 0.11	0.41 ± 0.15
Nonrigid Registration[8]	0.68 ± 0.08	0.58 ± 0.10
ICM Registration in [6]	0.76 ± 0.04	0.78 ± 0.06
Proposed Method	0.79 ± 0.04	0.82 ± 0.05

Table 2. Evaluation of Registration Error: PFP (%)

	Bladder	Uterus
Rigid Registration	33.57 ± 5.34	30.92 ± 5.22
Nonrigid Registration[8]	19.21 ± 3.41	21.78 ± 4.03
ICM Registration in [6]	9.36 ± 1.32	11.63 ± 1.80
Proposed Method	9.04 ± 2.11	10.54 ± 1.75

From Table 1 and Table 2, we found that the rigid registration performed the poorest out of all the registrations algorithms. While the nonrigid registration and ICM based registration performed much better, however, these techniques did not take the presence and regression of tumor into account. Since the tumor (Gross Tumor Volume,GTV) is adjacent to the bladder and uterus, mis-match of the tumor may affect the alignment of bladder or uterus. Table 1 and Table 2 show that the proposed method outperformed these registration methods at aligning soft organs.

3.2 Detection Results

Fig.3 presents the surface of the 3D detected tumor using the proposed method.

We compare the tumor images obtained using our method with the manual detection performed by a clinician. For the proposed method, we set a threshold for the tumor probability map M . Fig.4(a)-(j) provide the comparison between the proposed method and the detection by an expert. The tumor binary images are obtained with the probability threshold 0.7, which presents all the voxels that have have an over 70% chance to be in the tumor. From our experiments, we have found that the detection results are not sensitive to the threshold. The thresholds between 0.5 and 0.85 give almost the same detection results.

From the experiments, we find that the tumor shape has a strong influence on the performance of the detection results. As a global trend, algorithms show more accurate detection on well-defined (e.g. Fig.4(a)(b)) than on ill-defined masses (e.g. Fig.4(c)(d)). The tumor size has a weak influence on the detection performance. There is not a specific size where the algorithm performs poorly, i.e. the algorithm is not sensitive to the tumor size, as shown in Fig.4.

Using the expert's manual segmentation as ground truth, we quantitatively evaluate our detection results in Table 3, which shows a consistent agreement of our detection with the expert's delineation.

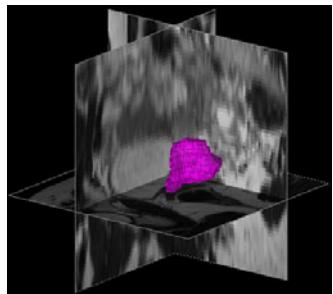


Fig. 3. 3D surface of the detected tumor using proposed method

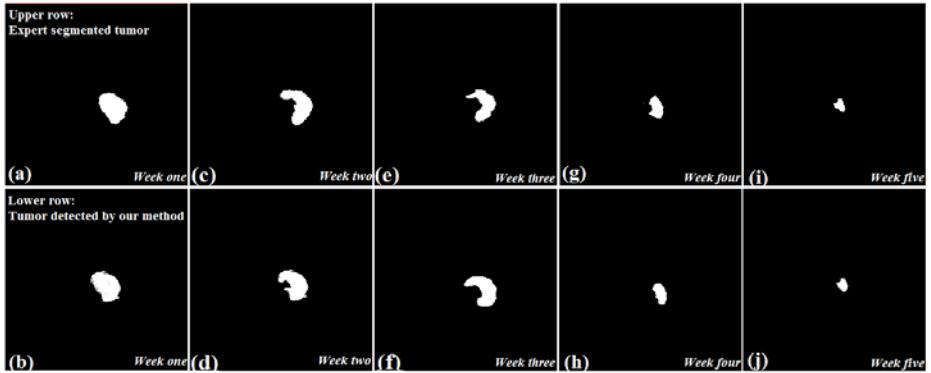


Fig. 4. Detection of cervical tumor for a patient. (a)-(j): Comparison. Top: Tumor outlined manually by a clinician. Bottom: Detection results using our proposed algorithm. Results from course of treatment (five weeks) are shown here.

Table 3. Evaluation of Detection: Tumor Overlaps (%)

	Week 1	Week 2	Week 3	Week 4	Week 5
Overlap	83.22 ± 5.34	78.81 ± 5.15	84.68 ± 5.34	80.57 ± 4.88	81.76 ± 6.93

4 Discussion and Conclusion

In this paper, a Bayesian framework for nonrigid registration and abnormality detection has been presented. The nonrigid registration part matches intensity information while taking the tumor estimation into consideration. We define the intensity matching as a mixture of two distributions which describe statistically image gray-level variations for both pixel classes (i.e. tumor class and normal tissue class). These mixture distributions are weighted by the tumor detection map which assigns to each voxel its probability of abnormality. It is shown on clinical data of patients with cervical cancer that the proposed method outperforms the previous image registration approaches using intensity information. In the future, we plan to develop a system that incorporates a physical tumor regression model and a shape prediction module, it is easy to calculate the location changes of the tumor for diagnosis and assessment, so that we can precisely guide the interventional devices toward the tumor during radiation therapy.

Acknowledgements. This work is supported by NIH/NIBIB Grant R01EB002164. We would like to thank Dr. Michael F. Milosevic and Dr. David A. Jaffray from Princess Margaret Hospital for providing the data and manual delineations.

References

1. Nag, S., Chao, C., Martinez, A., Thomadsen, B.: The american brachytherapy society recommendations for low-dose-rate brachytherapy for carcinoma of the cervix. *Int. J. Radiation Oncology Biology Physics* 52(1), 33–48 (2002)
2. van de Bunt, L., van de Heide, U.A., Ketelaars, M., de Kort, G.A.P., Jurgenliemk-Schulz, I.M.: Conventional conformal, and intensity-modulated radiation therapy treatment planning of external beam radiotherapy for cervical cancer: The impact of tumor regression. *International Journal of Radiation Oncology, Biology, Physics* 64(1), 189–196 (2006)
3. Jaffray, D.A., Carbone, M., Menard, C., Breen, S.: Image-guided radiation therapy: Emergence of MR-guided radiation treatment (MRgRT) systems. In: *Medical Imaging 2010: Physics of Medical Imaging*, vol. 7622, pp. 1–12 (2010)
4. Staring, M., van der Heide, U.A., Klein, S., Viergever, M., Pluim, J.: Registration of cervical MRI using multifeature mutual information. *IEEE Transactions on Medical Imaging* 28(9), 1412 (2009)
5. Lu, C., Chelikani, S., Papademetris, X., Knisely, J.P., Milosevic, M.F., Chen, Z., Jaffray, D.A., Staib, L.H., Duncan, J.S.: An integrated approach to segmentation and nonrigid registration for application in image-guided pelvic radiotherapy. *Medical Image Analysis* 15(5) (2011), doi:10.1016/j.media.2011.05.010
6. Lu, C., Chelikani, S., Chen, Z., Papademetris, X., Staib, L.H., Duncan, J.S.: Integrated segmentation and nonrigid registration for application in prostate image-guided radiotherapy. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) *MICCAI 2010. LNCS*, vol. 6361, pp. 53–60. Springer, Heidelberg (2010)
7. Oliver, A., Freixenet, J., Martí, J., Perez, E., Pont, J., Denton, E.R.E., Zwiggelaar, R.: A review of automatic mass detection and segmentation in mammographic images. *Medical Image Analysis* 14(2), 87 (2010)
8. Rueckert, D., Sonoda, L.I., Hayes, C., Hill, D.L.G., Leach, M.O., Hawkes, D.J.: Nonrigid registration using free-form deformations: Application to breast MR images. *IEEE Trans. Med. Imag.* 18(8), 712–721 (1999)
9. Greene, W.H., Chelikani, S., Purushothaman, K., Chen, Z., Papademetris, X., Staib, L.H., Duncan, J.S.: Constrained non-rigid registration for use in image-guided adaptive radiotherapy. *Medical Image Analysis* 13(5), 809–817 (2009)
10. Lu, C., Chelikani, S., Duncan, J.S.: A unified framework for joint segmentation, nonrigid registration and tumor detection: Application to MR-guided radiotherapy. In: Székely, G., Hahn, H.K. (eds.) *IPMI 2011. LNCS*, vol. 6801, pp. 525–537. Springer, Heidelberg (2011)
11. Lu, S.C., Chelikani, Papademetris, X., Staib, L., Duncan, J.: Constrained non-rigid registration using Lagrange multipliers for application in prostate radiotherapy. In: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2010)*, pp. 133–138 (June 2010)
12. Hamm, B., Forstner, R.: Section 3.2.1. General MR Appearance. In: *MRI and CT of the Female Pelvis*, p. 139 (2007)

Segmentation Based Features for Lymph Node Detection from 3-D Chest CT

Johannes Feulner^{1,3}, S. Kevin Zhou², Matthias Hammon⁴,
Joachim Hornegger¹, and Dorin Comaniciu²

¹ Pattern Recognition Lab, University of Erlangen-Nuremberg, Germany

² Siemens Corporate Research, Princeton, NJ, USA

³ Siemens Corporate Technology, Erlangen, Germany

⁴ Radiology Institute, University Hospital Erlangen, Germany

Abstract. Lymph nodes routinely need to be considered in clinical practice in all kinds of oncological examinations. Automatic detection of lymph nodes from chest CT data is however challenging because of low contrast and clutter. Sliding window detectors using traditional features easily get confused by similar structures like muscles and vessels. It recently has been proposed to combine segmentation and detection to improve the detection performance. Features extracted from a segmentation that is initialized with a detection candidate can be used to train a classifier that decides whether the detection is a true or false positive. In this paper, the graph cuts method is adapted to the problem of lymph nodes segmentation. We propose a setting that requires only a single positive seed and at the same time solves the small cut problem of graph cuts. Furthermore, we propose a feature set that is extracted from the candidate segmentation. A classifier is trained on this feature set and used to reject false alarms. Cross validation on 54 CT datasets showed that the proposed system reaches a detection rate of 60.9% with only 6.1 false alarms per volume image, which is better than the current state of the art of mediastinal lymph node detection.

1 Introduction

In clinical practice, radiologists commonly have to consider lymph nodes, especially in the area of the mediastinum (the area between the lungs). They are highly relevant in case of cancer [4,10]. Affected lymph nodes are typically enlarged. The total volume of all nodes, the number of nodes, the spatial distribution or changes over time can give a physician important information about the progress of the disease and the effectiveness of the treatment. Physicians typically use CT for the assessment. Computing such statistics requires to detect and/or segment the lymph nodes. Finding and measuring lymph nodes manually is however very time consuming and therefore not done in clinical practice. Furthermore, there is a high inter and even intra observer variability [6].

An automatic system that detects and segments lymph nodes from CT data would therefore be of high clinical use. It however has to cope with clutter and low contrast because lymph nodes are often in the neighborhood of muscles and

vessels and have at the same time a similar attenuation coefficient. Furthermore, there is a great variability in both the size and the shape of lymph nodes.

The topic has received increasing attention in the last five years. In [9,5], lymph nodes are detected by a cascade of filters (Hessian based, morphological operations and a so-called 3-D Min-DD filter). In [3], lymph nodes are detected and segmented by fitting a mass-spring model at different positions.

In this paper, we follow [6,1] that proposed two data driven approaches. As in [6], a discriminative model of the lymph node appearance is combined with a spatial prior probability that contains anatomical knowledge about where lymph nodes are likely to appear. Similar to [1], detection and segmentation is combined to improve the detection performance. We introduce a feature set that is extracted from a candidate segmentation. A classifier is trained on the feature set to decide whether a segmentation is an actual lymph node or a false positive detection.

In contrast to [1], we do not fit a sphere to a lymph node. Instead, we adapt graph cuts to the problem of lymph node segmentation. Prior knowledge of the lymph node appearance is included by selecting the weights of the graph according to manually annotated data. The segmentation is initialized with a single point from the detection result. To overcome a major problem of graph cuts segmentation, the small cut problem, we introduce an additional radial weighting of the graph that is well suited for segmenting blob-like structures.

Fig. 1 shows an overview of our system that consists of four stages. In stages 1–3, a list of possible lymph node center positions is generated. Stage 4 is the major contribution of this paper: Here, the detected lymph nodes are verified using features extracted from a candidate segmentation. At all stages, the detection score is weighted using a spatial prior probability as introduced in [6].

The remainder of this paper is structured as follows: Section 2 presents our method that jointly detects and segments lymph nodes, section 3 presents experiments and results, and section 4 concludes the paper.

2 Proposed Method for Detecting and Segmenting Lymph Nodes

2.1 Candidate Generation

In each of the first three stages, a binary classifier is trained to learn the probability $p(m = 1|t)$ of observing a true lymph node at position t . Here, m is the binary class variable. In stages one and two, we follow the approach of [6]: A probabilistic boosting tree (PBT) classifier [13] is trained with 3-D Haar-like features. A PBT is a binary decision tree with a strong Ada-Boost classifier at each node. Haar-like features are box features that are simple but powerful because they can be computed very efficiently. At stages two to four, only the detections

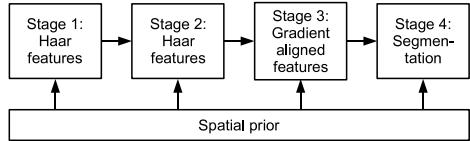


Fig. 1. Overview of the detection pipeline

from the previous stage are considered during test. In the training phase, negative examples of all stages except the first one are generated from false alarms of the previous one. Thus, the classifiers get specialized on the hard examples. The third classifier uses gradient-aligned features (GAF) introduced in [1]. The idea is to extract the features at locations of gradient jumps, because these are hints for the object boundary. Starting from the lymph node center candidate \mathbf{t} , gradients are computed along rays in radial direction. At local maxima of the gradient magnitude, simple point features are computed.

Lymph nodes cannot appear anywhere, they always lie in fat tissue. They are however not homogeneously distributed there but much more common in certain regions. As introduced in [6], we use a spatial prior probability of lymphatic tissue to model this anatomical knowledge. It can be thought of a probabilistic atlas that is learned from manually annotated data. It is registered to a test image using a set of anatomical landmarks that are detected automatically. Further details about this spatial prior can be found in [6].

2.2 Joint Detection and Segmentation

Segmenting node-like structures using graph cuts. At this point, we already have the center \mathbf{t} of a detected lymph node candidate from our previous detection steps one to three. We consider a sub-image cropped from the original volume image such that \mathbf{t} is centered in the sub image. The size of the sub-image remains fixed at $4 \times 4 \times 4\text{cm}$. This is relatively large and ensures that almost all lymph nodes fit into this window. Apart from knowing the center, we also know the intensity distribution of lymph nodes, of the background, and the 2-D joint distribution of voxel pairs on the object boundary from manually segmented data. We furthermore know that lymph nodes have a blob-like shape. In the sub-image, the lymph node is segmented using the graph cuts method for seeded image segmentation [8,2]. The voxels form the vertices of the graph, and neighboring voxels are connected. We propose a setting that incorporates all the prior knowledge mentioned above.

It was shown in [8] that the energy function optimized by graph cuts is

$$\hat{\mathbf{x}} = \underset{\mathbf{x}}{\operatorname{argmax}} \sum_i \lambda_i x_i + \frac{1}{2} \sum_{ij} \beta_{ij} \delta(x_i, x_j). \quad (1)$$

Here, $x_i \in \{0, 1\}$ is the binary label of voxel i that is one for “foreground” or zero for “background”, $\hat{\mathbf{x}} = (\hat{x}_1, \dots, \hat{x}_N)$ is the vector of all labels, $\delta(a, b)$ is the Kronecker delta, λ_i is the unary weight of voxel i , β_{ij} is the binary weight (or capacity) of the (directed) edge from voxel i to voxel j , and N is the number of voxels in the sub-image. A high β_{ij} value reflects that voxels i and j are likely to have the same label. A high λ_i value means that, without knowing anything about its neighborhood, voxel i is more likely to be foreground.

Since the center \mathbf{t} of the sub-image is assumed to be the center of the lymph node, it is used as positive seed and its λ_{it} value is set to ∞ . The boundary voxels of the sub-image are marked as negative seeds and their unary weights are set to $-\infty$.

Graph cut is however known to be sensitive to the number of seeds [11]. If all other unary capacities λ_i were set to zero, and all binary capacities β_{ij} to some positive constant, then the smallest cut that separates the source from the sink would simply separate the positive seed from its direct neighbors. This is also known as the small-cut problem of graph cuts. In this setting, the problem can be solved by simply adding a factor $\frac{1}{r_{ij}^2}$ to the capacities β_{ij} , were r_{ij} denotes the distance of the center point of the edge from voxel i and voxel j to the positive seed at t . If the original capacities β'_{ij} are constant and $\beta_{ij} = \frac{1}{r_{ij}^2} \beta'_{ij}$, then the integrated capacity $B(r)$

$$B(r) = \sum_{(i,j) \in \mathcal{B}(r)} \beta_{ij} \quad (2)$$

over a sphere centered at t is nearly constant for different radii r (it is not exactly constant because of discrete voxels). In (2), $\mathcal{B}(r)$ denotes the set of edges intersected by the sphere centered at t with radius r . Now there is no bias any more toward a small cut. Because of the smaller surface, spherical cuts are preferred over non-spherical cuts, which is a desirable property for the purpose of segmenting a node-like structure. This method is not only simple, is also comes at no additional computational costs.

Other shape priors have been proposed for graph cuts segmentation. In [12], a prior for elliptic shapes was introduced. However, the segmentation must be solved iteratively. In [7], a method that favors cuts that are orthogonal to the line from the current point to the center was proposed. This is effectively a prior for blob-like structures but does not solve the small cut problem. A prior for star-shaped structures and also a balloon force that corresponds to a certain boundary length was introduced in [14]. This solves the small cut problem, but the balloon force is optimized iteratively.

Most lymph nodes have an approximately constant attenuation coefficient. This allows selecting the graph capacities according to intensity histograms. We set the unary capacity λ_i to

$$\lambda_i = \log \frac{p^u(\text{FG}|I_i)}{1 - p^u(\text{FG}|I_i)} \quad (3)$$

the logarithm of the odds that voxel i is foreground (FG) given its intensity value I_i . The probability $p(\text{FG}|I_i)$ is estimated non-parametrically using a histogram. u is a normalizing constant that is used to balance the influence of the unary and binary capacities. It was set to 0.13 in the experiments. The binary capacity β_{ij} is set to

$$\beta_{ij} = -\frac{\log [p(\text{out}_{ij})p(B|I_i, I_j)]}{r_{ij}^2 \text{dist}(i, j)} \quad \text{with} \quad p(\text{out}_{ij}) = \frac{\cos \alpha_{ij} + 1}{2}. \quad (4)$$

Here, $p(B|I_i, I_j)$ denotes the probability of observing the object boundary (B) between the adjacent voxels i and j given the intensity I_i of the voxel inside and I_j of the voxel that it assumed to be outside the segmentation. Note that this is not symmetric. $\text{dist}(i, j)$ denotes the euclidean length of the edge from voxel i to voxel j , and $p(\text{out}_{ij})$ is the estimated probability that the edge from i to j is

pointing in outward direction. Here, α_{ij} is the angle between the edge from i to j and the line from the positive seed to the center of the edge. Thus, $\cos \alpha_{ij} = 1$ if the edge is pointing away from the central seed, and $\cos \alpha_{ij} = -1$ if it is pointing toward the center. By allowing directed edge capacities, we incorporate additional knowledge about the boundary appearance. The term $p(B|I_i, I_j)$ in (4) can be expressed as

$$p(B|I_i, I_j) = \frac{p(B, I_i, I_j)}{p(I_i, I_j)}. \quad (5)$$

Both $p(B, I_i, I_j)$ and $p(I_i, I_j)$ are estimated non-parametrically using joint intensity histograms. However, $p(B, I_i, I_j)$ is sparse because of a limited number of training examples of points on the boundary of lymph nodes. Therefore, $p(B|I_i, I_j)$ is smoothed with a Gaussian filter with $\sigma = 40\text{HU}$.

Segmentation based features. As final stage in the detection cascade, an Ada-Boost classifier is trained with features extracted from the segmentation that was initialized with the detected lymph node center t to learn whether t is a true lymph node or a false detection.

The first kind of features are histogram based: A hierarchy of normalized histograms of the intensity values inside the segmentation is computed. The histogram at the first level has 256 bins. Each bin is one Hounsfield unit wide, and the first bin corresponds to -128 HU. Lymph nodes typically fall into this range of HU values. At the next level, the number of bins is halved, and the width of each bin is doubled. In total, seven levels are used. The entry of each bin of each pyramid level is a scalar feature. The second kind of features are again based on a hierarchy of histograms, but the histograms are now computed from the 3mm wide neighborhood of the segmentation. Additionally, we use the second, third and fourth central moments of the histograms both inside and outside the segmentation. Next, 100 points are randomly sampled from the surface of the segmentation. As proposed in [1], the points are sorted by their gradient magnitude to enumerate them. The surface normal at each point is sampled at seven positions with a spacing of 1mm between the samples. At each sample, simple point features are computed. All scalar features at all samples at all normals at all points are added to the feature pool. Finally, the volume, the surface, the sphericity, the maximum flow value and the maximum flow divided by the surface are used. In total, the feature pool contains 51436 features. During training, AdaBoost selects 270 of them.

3 Results

The proposed method has been evaluated on 54 CT datasets showing the chest area of lymphoma patients. The voxel spacing typically was $0.8 \times 0.8 \times 1\text{mm}^3$. All datasets were resampled to an isotropic $1 \times 1 \times 1\text{mm}^3$ resolution. The mediastinal lymph nodes were manually segmented, and the segmentations were reviewed by an experienced radiologist.

The detection performance was evaluated using threefold cross-validation. The classifiers were only trained on lymph nodes that have a minimum size of 10mm

in at least two dimensions. Smaller lymph nodes are usually not pathologic [10] and were therefore neglected. Among the segmented lymph nodes, 289 were used for training. To achieve a better generalization and to avoid overfitting, the training data was mirrored by all three coordinate planes, resulting in $2^3 = 8$ times more training examples. For testing, only the original data was used. In the testing phase, a lymph node is considered as detected if the center t of a detection is inside the tight axis-aligned bounding box of the lymph node. A lymph node is considered as false negative (FN) if its size is at least 10mm and it is not detected.

Occasionally, two or more detections are close together. In order to reduce the number of such double detections, the detected centers are spatially clustered and merged. Two detections are merged if their distance is 6mm or less. The confidence value of the merged detection is set to the sum of the original ones.

Fig.2 shows the result of cross-validation as FROC (free-response receiver operating characteristic) curves. The final verification step reduces the number of false positives (FP) considerably and improves the true positive rate (TPR) by

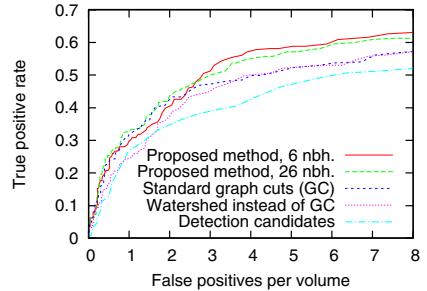


Fig. 2. FROC curve of the detection performance

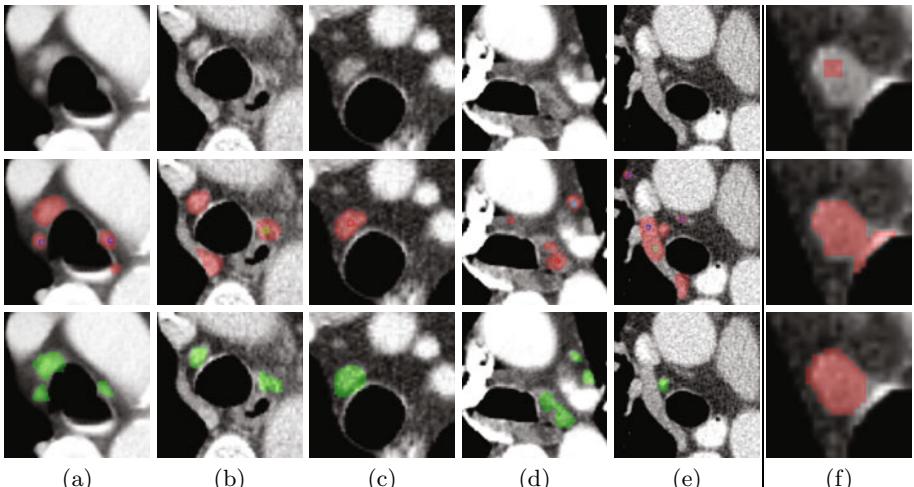


Fig. 3. (a-e): Detection and segmentation examples shown in 2-D. Top row: Plain CT slices. Second row: Detections (colored boxes) and resulting segmentations (red). The detection score is color coded in HSV color space. Violet means lowest, red means highest score. Bottom row: Manual ground truth segmentations. (f): Manually initialized segmentations with different binary edge capacities. See text for details.

Table 1. Detection results compared to state of the art methods

Method	Body region	num.	vol. size/mm	TP	FP	FN	TPR	FP per vol.
Kitasaka et al. [9]	Abdomen	5	> 5.0	126	290	95	57.0%	58
Feuerstein et al. [5]	Mediastinum	5	> 1.5	87	567	19	82.1%	113
Dornheim [3]	Neck	1	> 8.0	29	9	0	100%	9
Feulner et al. [6]	Mediastinum	54	>10.0	139	379	127	52.3%	7.0
Barbu et al. [1]	Axillary	101	>10.0	298	101	64	82.3%	1.0
This method	Mediastinum	54	>10.0	153	167	136	52.9%	3.1
This method	Mediastinum	54	>10.0	176	332	113	60.9%	6.1

38% at 3.5 FP per volume and by 20% at 7 FP per volume (red and cyan curve). Using a 26 or a six neighborhood system in the graph cuts segmentation step does not significantly affect the detection performance (red and green curve). We however noticed that a 26 neighborhood produces smoother segmentations. We also exchanged our segmentation method with either standard graph cuts with weights $\lambda_i = 0$, $\beta_{ij} = \exp(-(I_i - I_j)^2/2\sigma_\beta^2)$ ($\sigma_\beta = 16\text{HU}$) that are popular in the literature [2] or a watershed segmentation and measured the detection accuracy. The proposed segmentation method reaches a higher recall.

Example segmentations and detections on unseen data are shown in Fig. 3 (a-e). The bigger lymph nodes are detected. There are however some false alarms, especially on vessels (e). Fig. 3 (f) shows manually initialized segmentations with different edge capacities. The top and center image were segmented with standard graph cuts weights. This often causes small cuts (top image, $\sigma_\beta = 32\text{HU}$) or leakage and rugged segmentations (center image, $\sigma_\beta = 16\text{HU}$). This is solved by using the weights proposed in this paper (bottom image).

In Table 1, other methods are listed together with their performance for comparison. The comparability is however limited because of different data, different criterions for a detection, different body regions and different minimum lymph node sizes used for evaluation. Both [9] and [5] report a very high number of false alarms and also consider a lymph node already as detected if there is just overlap with the automatic segmentation. In [3], very good results are reported, but the method was evaluated on a single dataset. In [1], good results are reported for the axillary region. Lymph nodes in the axillary regions are however easier to detect because they are mostly isolated in fat tissue and less surrounded by clutter as in the mediastinal region. The method of [6] was evaluated on the mediastinum. While [6] reaches a detection rate of 52.3% at 7.0 false alarms per volume, this method detects 52.9% of the lymph nodes with only 3.1 false alarms per volume and 60.9% with 6.1 false alarms per volume.

In [6], the intra-observer variability (IOV) of mediastinal lymph node detection is reported to have a TPR of 54.8% at 0.8 FP per volume. Even though 0.8 FP per volume is a very good value, it demonstrates that finding mediastinal lymph nodes is very challenging also for humans.

The computational requirements of the proposed methods were measured on a standard dual core PC with 2.2GHz. In total, detecting and segmenting the mediastinal lymph nodes in a CT volume takes 60.2s if a six neighborhood system

is used and 101.9s with a 26 neighborhood system. Computing the spatial prior takes 19.5s and is already included.

4 Conclusion

The contribution of this paper is twofold: First, we propose using a single centered positive seed for graph cuts and a radial weighting of the edge capacities as a segmentation method for blob-like structures that is not prone to the small cut problem. Second, a feature set is proposed that is extracted from a segmentation. A classifier is trained on the feature set and rejects detections with poor segmentation results. The segmentation based verification step clearly helps to detect mediastinal lymph nodes. Our proposed system reaches a detection rate of 60.9% at 6.1 false alarms per volume. This is better than the state of the art in mediastinal lymph node detection [6]. At the moment, there are especially false alarms on vessels. Combining the proposed method with a good vessel detector should further improve the detection performance.

References

1. Barbu, A., Suehling, M., Xu, X., Liu, D., Zhou, S.K., Comaniciu, D.: Automatic detection and segmentation of axillary lymph nodes. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6361, pp. 28–36. Springer, Heidelberg (2010)
2. Boykov, Y., Funka-Lea, G.: Graph cuts and efficient n-d image segmentation. IJCV 70, 109–131 (2006), <http://dx.doi.org/10.1007/s11263-006-7934-5>
3. Dornheim, L., Dornheim, J.: Automatische detektion von lymphknoten in ct-datensätzen des halses. In: Bildverarbeitung für die Medizin, pp. 308–312 (2008)
4. Duwe, B.V., Sterman, D.H., Musani, A.I.: Tumors of the mediastinum. Chest 128(4), 2893–2909 (2005)
5. Feuerstein, M., Deguchi, D., Kitasaka, T., Iwano, S., Imaizumi, K., Hasegawa, Y., Suenaga, Y., Mori, K.: Automatic mediastinal lymph node detection in chest ct. In: SPIE Medical Imaging, Orlando, Florida, USA (February 2009)
6. Feulner, J., Zhou, S.K., Huber, M., Hornegger, J., Comaniciu, D., Cavallaro, A.: Lymph node detection in 3-d chest ct using a spatial prior probability. In: CVPR (2010)
7. Funka-lea, G., Boykov, Y., Florin, C., Jolly, M.-p., Moreau-gobard, R., Ramaraj, R., Rinck, D.: Automatic heart isolation for ct coronary visualization using graph-cuts. In: IEEE International Symposium on Biomedical Imaging (2006)
8. Greig, D.M., Porteous, B.T., Seheult, A.H.: Exact maximum a posteriori estimation for binary images. JRSS Series B 51(2), 271–279 (1989)
9. Kitasaka, T., Tsujimura, Y., Nakamura, Y., Mori, K., Suenaga, Y., Ito, M., Nawano, S.: Automated extraction of lymph nodes from 3-D abdominal CT images using 3-D minimum directional difference filter. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part II. LNCS, vol. 4792, pp. 336–343. Springer, Heidelberg (2007)
10. de Langen, A.J., Raijmakers, P., Riphagen, I., Paul, M.A., Hoekstra, O.S.: The size of mediastinal lymph nodes and its relation with metastatic involvement: a meta-analysis. Eur. J. Cardiothorac. Surg. 29(1), 26–29 (2006)

11. Sinop, A., Grady, L.: A seeded image segmentation framework unifying graph cuts and random walker which yields a new algorithm. In: IEEE 11th International Conference on Computer Vision, ICCV 2007, pp. 1–8 (2007)
12. Slabaugh, G., Unal, G.: Graph cuts segmentation using an elliptical shape prior. In: ICIP, vol. 2, pp. II–1222–II–1225 (2005)
13. Tu, Z.: Probabilistic boosting-tree: learning discriminative models for classification, recognition, and clustering. In: ICCV, vol. 2, pp. 1589–1596 (2005)
14. Veksler, O.: Star shape prior for graph-cut image segmentation. In: Forsyth, D., Torr, P., Zisserman, A. (eds.) ECCV 2008, Part III. LNCS, vol. 5304, pp. 454–467. Springer, Heidelberg (2008)

Segmenting Hippocampus from 7.0 Tesla MR Images by Combining Multiple Atlases and Auto-Context Models

Minjeong Kim¹, Guorong Wu¹, Wei Li¹, Li Wang¹, Young-Don Son²,
Zang-Hee Cho², and Dinggang Shen¹

¹ Department of Radiology and BRIC, University of North Carolina at Chapel Hill

² Neuroscience Research Institute, Gachon University of Medicine and Science, Incheon, Korea

Abstract. In investigation of neurological diseases, accurate measurement of hippocampus is very important for differentiating inter-subject difference and subtle longitudinal change. Although many automatic segmentation methods have been developed, their performance can be limited by the poor image contrast of hippocampus in the MR images, acquired from either 1.5T or 3.0T scanner. Recently, the emergence of 7.0T scanner sheds new light on the study of hippocampus by providing much higher contrast and resolution. But the automatic segmentation algorithm for 7.0T images still lags behind the development of high-resolution imaging techniques. In this paper, we present a learning-based algorithm for segmenting hippocampi from 7.0T images, by using multi-atlases technique and auto-context models. Specifically, for each atlas (along with other aligned atlases), Auto-Context Model (ACM) is performed to iteratively construct a sequence of classifiers by integrating both image appearance and context features in the local patch. Since there exist plenty of texture information in 7.0T images, more advanced texture features are also extracted and incorporated into the ACM during the training stage. With the use of multiple atlases, multiple sequences of ACM-based classifiers will be trained, respectively in each atlas' space. Thus, in the application stage, a new image will be segmented by first applying the sequence of the learned classifiers of each atlas to it, and then fusing multiple segmentation results from multiple atlases (or multiple sequences of classifiers) by a label-fusion technique. Experimental results on the six 7.0T images with voxel size of $0.35 \times 0.35 \times 0.35 mm^3$ show much better results obtained by our method than by the method using only the conventional auto-context model.

1 Introduction

Numerous efforts have been made for developing hippocampus segmentation methods in the field of medical image processing [1-6], since accurate measurement of hippocampus is significant for study of neurological diseases, i.e., Alzheimer's disease. However, due to the tiny size of hippocampus ($\approx 35 \times 15 \times 7 mm^3$) and also the complexity of surrounding structures, the accuracy of hippocampus segmentation is limited by the low imaging resolution (often with voxel size of $1 \times 1 \times 1 mm^3$).

The learning-based methods [7, 8] have been proven powerful for hippocampus segmentation. Recently, auto-context model (ACM) has been proposed in [9] to automatically segment hippocampus by constructing the classifier based on not only

the image appearance, but also the context features (i.e., a map of classification confidence) in the local patch. In each iteration of ACM, the context information is updated according to the latest classification results and then used to guide the training of the next classifier by Adaboost [10] in the next round of training. To be successful, a large number of image and context features are generally extracted in ACM, which includes intensity, position, and neighborhood features (such as intensity mean and variance, and Haar features at different scales). Although the ACM is effective by integrating the low-level appearance features with the high-level context and implicit shape information, its performance in hippocampus segmentation could still be limited by the distinctiveness of features, which is directly related with imaging resolution.

On the other hand, the high-resolution imaging technique has been rapidly developed after the 3.0T scanner becomes popular in the modern neurological study. It has been shown in [11] that hippocampi can be better manually labeled on 7.0T images than the low-resolution 3.0T images. Thus, researchers have started to apply the previous segmentation algorithms, developed for 1.5T or 3.0T images, to 7.0T images. However, the performance is limited, mainly because the features in 7.0T are much richer and significantly different from those in 1.5T or 3.0T. To demonstrate the rich texture and high enhancement on image contrast in 7.0T image, we show one typical slice with hippocampi (with the resolution of $0.35 \times 0.35 \times 0.35 \text{ mm}^3$) in Fig. 1 (b), along with a similar slice scanned by 1.5T scanner (with the resolution of $1 \times 1 \times 1 \text{ mm}^3$) in Fig. 1 (a).

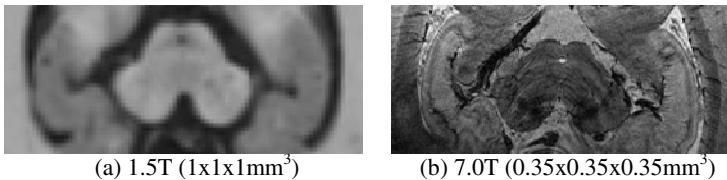


Fig. 1. Large difference between 1.5T (a) and 7.0T (b) MR images

To fully utilize the rich features provided by 7.0T images, we present a new learning-based hippocampus segmentation algorithm, using multi-atlases technique and auto-context models. In particular, ACM is performed to train the classifiers for each atlas and its aligned atlases, based on their respective manually-labeled hippocampi. Importantly, since 7.0T images provide texture information, more advanced texture features are also incorporated into the ACM during the training stage. To segment the hippocampi on a new image, all atlases, along with their respective classifiers learned in the training stage, will be first transformed onto the new image by a registration technique. Then each point in the new image will be classified by the respective classifiers of each atlas, and its final segmentation result will be obtained by fusing all results from all atlases.

Our learning-based hippocampus segmentation method has been evaluated extensively on 7.0T images. It can produce 7%~12% more overlap ratio than the method using only the conventional ACM, because of using multi-atlases technique and also the incorporation of more advanced texture information into the ACM framework in our method.

2 Method

The goal of our learning-based hippocampus segmentation algorithm is to accurately label each point $x \in \Omega$ in a new subject into either positive (i.e., hippocampus) or negative (i.e., non-hippocampus). Fig. 2 shows the overview of our proposed method. It basically includes two stages, i.e., training stage and test stage, as shown respectively by the blue and red arrows/block in Fig. 2. In the training stage, totally M images $\mathbf{I} = \{I_i(x)|x \in \Omega, i = 1, \dots, M\}$ and their corresponding manual hippocampi labels $\mathbf{L} = \{L_i(x)|x \in \Omega, i = 1, \dots, M\}$ are used as atlases for training the ACM-based classifiers. For each atlas I_i , the classifiers for hippocampus segmentation will be trained based on not only the current atlas I_i , but also other ($M - 1$) atlases that have been aligned onto the space of I_i . During the training, manual labels \mathbf{L} will be used as the ground truth, and the ACM technique will be used to construct the classifiers by using the local-patch appearance and context features, as described in Section 2.1. Note that we will finally obtain M sets of classifiers, corresponding to the M atlases. In the application stage, the segmentation of a test image will be conducted by applying the classifiers of each atlas to obtain a segmentation result, and then the multiple results from multiple atlases are fused into a final single result by the label-fusion technique, as detailed in Section 2.2.

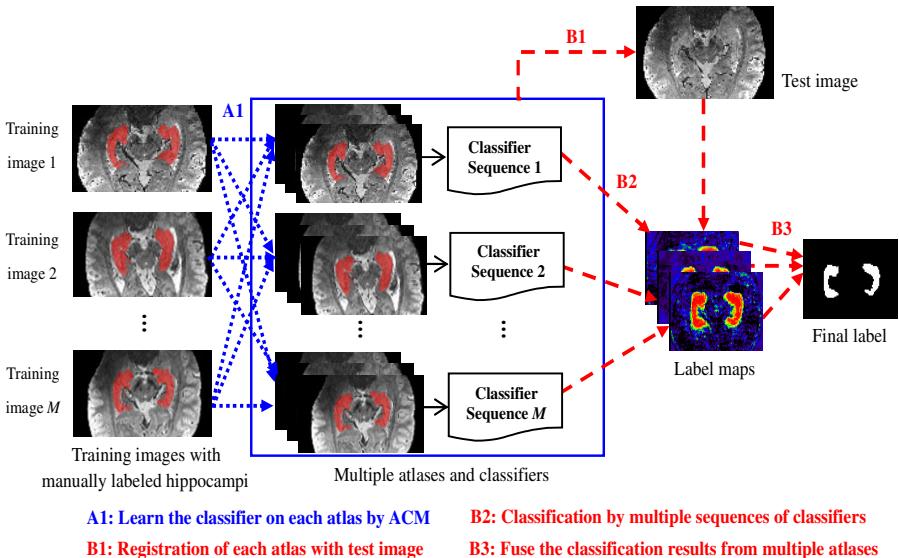


Fig. 2. Schematic illustration of the proposed hippocampus segmentation framework, which includes a training stage (blue arrows and block) and a testing stage (red arrows)

2.1 Learn the Classifiers in Each Atlas Space by Auto-Context Model

In the training stage, ACM is performed in the space of each atlas I_i to train a specific set of classifiers, from both I_i and other atlases I_j . Note that I_j is used here

to denote any atlas other than I_i . Before constructing the classifiers for each atlas I_i , all I_j s need to be aligned with I_i , which can be achieved by affine registration of manually-segmented hippocampi in I_i and I_j . Since the spatial context feature on classification (or confidence) map has been incorporated in the ACM, which has implicitly taken the shape priors into consideration, the registration accuracy in our method is not the main factor for the classification performance, compared to the multi-atlases-based segmentation methods which typically rely on deformable registration.

Auto-Context Model: Here we follow the notation in [12] to define X_i as the vector of all spatial positions in the atlas I_i . Suppose that there are totally N points in I_i , and thus $X_i = (x_1^i, \dots, x_n^i, \dots, x_N^i)$, where x_n^i ($n = 1, \dots, N$) represents a spatial position in I_i . Each X_i comes with a ground-truth label vector $Y_i = (y_1^i, \dots, y_n^i, \dots, y_N^i)$, where $y_n^i \in \{1, \dots, k, \dots, K\}$ is the class label for the associated x_n^i . In our application, $K = 2$ since we only need to distinguish between hippocampus and non-hippocampus labels. $P^t(X_i) = (P^t(x_1^i), \dots, P^t(x_n^i), \dots, P^t(x_N^i))$ is the classification map in the t -th iteration ($t = 0, \dots, T$), where each element $P^t(x_n^i) = p(y_n^i = k|x_n^i)$ is the likelihood probability of assigning label k to the point x_n^i , which can be obtained from the classification results. It is worth noting that the initial classification map $P^0(X_i)$ is a binary vector, assigned from the union of all aligned manual-labeling hippocampi to each atlas I_i .

Thus, the training set is defined as $S^t = \{X_i, Y_i, P^t(X_i)\} \cup \{[X_j, Y_j, P^t(X_j)], j = 1, \dots, M - 1\}$, where the first part $\{X_i, Y_i, P^t(X_i)\}$ is from the current atlas I_i and the latter part from all other atlases I_j after affine registration. Two kinds of features can be extracted for training classifiers: 1) image features computed from the local patch, and 2) context information captured from the classification map at a large number of sites, as detailed below.

Appearance and Context Features: Similar to [12], the image *appearance features* we used include three kinds of information, i.e., intensity, spatial location, and neighborhood features (such as intensity mean and variance, and Haar features at different scales). To further utilize the texture information in 7.0T images, the gray level co-occurrence matrix (GLCM) is also calculated in each local patch of point x_n^i under consideration. Then, the total 14 statistics on the matrix, e.g., contrast, correlation, entropy, energy, and homogeneity, are considered as additional appearance features in our method. The *context features* are extracted from the classification map. Specifically, for each point x_n^i , a number of rays in equal-degree intervals are extended out from the current point and we sparsely sample the context locations on these rays. For each context location, the classification probability and the mean probability within 3×3 windows are included as context features. It is worth noting that both image appearance and context features are calculated in a spherical patch with the radius of 21.

Training of ACM: Given the underlying atlas I_i and all other aligned atlases I_j , along with their manual-segmented hippocampi labels, the initial classification map

can be computed by the union of all aligned manual-segmented hippocampi labels (\mathbf{L}), as mentioned above. Then, ACM is able to iteratively find the optimal classifiers by repeating the following steps for T times:

- Update the training set S^t (as defined above) by using the newly-obtained classification maps $P^t(X_i)$ and $P^t(X_j)$;
- Train the classifier by using both image appearance and context features extracted from the local patch.
- Use the trained classifier to assign the label to each point in I_i and all I_j s.

For each atlas I_i under consideration, the output of ACM is a sequence of trained classifiers, i.e., T classifiers corresponding to the total T iterations. The same training procedure will be applied to all other atlases. Thus, in the end of training, for given M atlases, we will obtain M sequences of trained classifiers, where each sequence corresponds to each of M atlases and have totally T trained classifiers, as shown in the blue box of Fig. 2. In the application stage, by registering each atlas onto the test image, we can apply its respective sequence of trained classifiers to segment the test image iteratively, as shown in Fig. 3. As we can see from Fig. 3, the classification map is updated with the iterations of the ACM algorithm, which can finally lead to the correct segmentation of hippocampus.

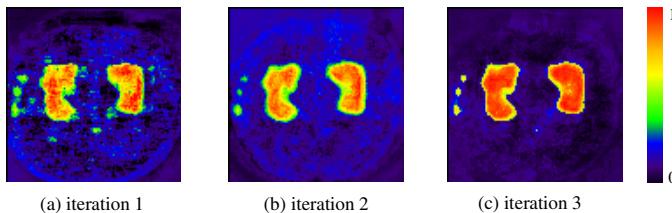


Fig. 3. The classification map of hippocampus at each iteration of the ACM algorithm

2.2 Multi-atlases Based Hippocampus Segmentation

In the application stage, hippocampus segmentation from a test image is completed by three steps (B1-B3), as displayed with red arrows in Fig. 2. In the first step (B1), all the atlases will be registered onto the test image, in order to map the classifiers learned in the training stage onto the test image space. In the second step (B2), the labeling of the test image is performed by the classifiers learned in each atlas space. Note that the procedure of labeling follows exactly the training procedure of ACM, by applying a sequence of the learned classifiers to compute the posterior marginal as shown in Fig. 3. Thus, a set of final classification maps, corresponding to the number of different atlases, will be obtained. In the third step (B3), those classification maps will be fused together by first applying the majority voting and then using a level-set segmentation algorithm to extract the final segmentation of hippocampus.

3 Experimental Results

In the following experiments, we evaluate the performance of our learning-based hippocampus segmentation method by six 7.0T MR images, each with the image size of $576 \times 512 \times 60$ and voxel resolution of $0.35 \times 0.35 \times 0.35 mm^3$. A leave-one-out test is used due to the limited number of samples. Specifically, at each leave-one-out case, one image is used as test image, and all other images are used as training images. In both training stage and testing stage, the affine registration is used to bring the images to the same space by the flirt algorithm in FSL library. We report both qualitative and quantitative results in the next.

3.1 Qualitative Results

In order for comprehensive evaluation, we compare the segmentation results by our method with those by the methods using only the conventional auto-context model (ACM), using the improved ACM with advanced texture features, and using the conventional ACM in the multi-atlases based framework. For fair comparison, we use the same parameters for the ACM used in the all methods, e.g., the feature sets and the number of iterations. Fig. 4 shows a segmented results on the selected test image by the four different segmentation methods (Figs. 4b, 4c, 4d, and 4e), along with manual segmentations (Fig. 4a). It can be observed that the segmented hippocampi by our method (Fig. 4e) are more similar to the ground-truth than those by the methods using only the conventional ACM (Fig. 4b), using the improved ACM (Fig. 4c), and using the conventional ACM in the multi-atlases based framework (Fig. 4d). The quantification results are provided in the next section.

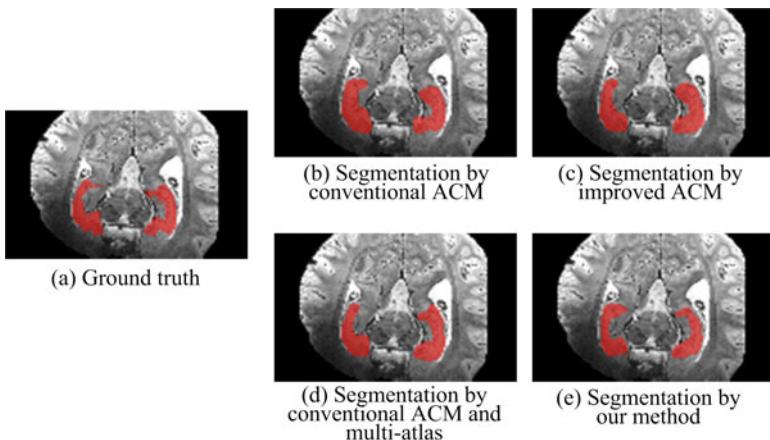


Fig. 4. Comparison of segmented hippocampus regions by (b) the method using only the conventional auto-context model, (c) the method using the improved ACM, (d) the method using the conventional ACM in the multi-atlases framework, and (e) our proposed method. Compared to the ground truth (a) obtained with manual labeling, our proposed method shows the best segmentation performance.

3.2 Quantitative Results

Due to the lack of enough number of 7.0T image samples, we performed a leave-one-out test to evaluate the segmentation performance of our method with the three other methods as described in Section 3.1. For each leave-one-out case, as mentioned above, we use one image for testing, and the rest 5 images for training the classifiers. Note that, in *the methods using only the auto-context model* (either conventional or improved one), we randomly select the space of one training sample as a common space, and warp all other training samples to this common space for building one sequence of classifiers. This sequence of classifiers will be applied to the test image for obtaining the final segmentation results. On the other hand, in *the methods using multi-atlases technique*, we regard each of the 5 training images as a common space and then warp all other training images to the common space for training one sequence of the classifiers. In this way, we will obtain 5 sequences of classifiers, corresponding to the 5 training samples which are selected as a common space, respectively. Then, these 5 sequences of classifiers can be applied to label all voxels in the test image. The final segmentation result for the test image can be obtained by fusing all results with a label-fusion scheme as described before.

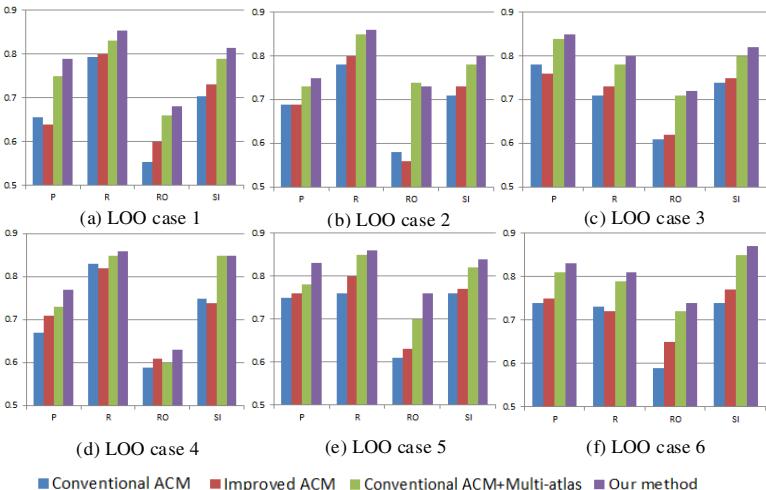


Fig. 5. Quantitative comparisons based on the 4 overlap metrics (precision (P), recall (R), relative overlap (RO), and similarity index (SI)) for the 6 leave-one-out (LOO) cases, which consistently shows the significant improvement in overlap scores by our method over other three methods (by using only the conventional ACM, improved ACM, and conventional ACM in the multi-atlases framework)

For quantitative comparison of the four methods, we use the following 4 overlap metrics: precision (P), recall (R), relative overlap (RO), and similarity index (SI), as defined below:

$$P = \frac{V(A \cap B)}{V(B)}, \quad R = \frac{V(A \cap B)}{V(A)}, \quad RO = \frac{V(A \cap B)}{V(A \cup B)} \quad \text{and} \quad SI = \frac{V(A \cap B)}{\{(V(A) + V(B))/2\}} \quad (1)$$

where $V(A)$ is the volume of the ground-truth segmentation such as manual segmentation, and $V(B)$ is the volume of automatic segmentation. Fig. 5 shows the results of 4 overlap metrics for each of 6 leave-one-out cases (LOO cases 1-6) by the four methods, indicating that our method consistently outperforms all other three methods. Especially, in the comparison with the method using only the conventional ACM, our method gains the improvement of 8.9%, 7.3%, 12.1% and 9.8% on average for the 4 overlap metrics, P , R , RO , and SI , respectively. On the other hand, the other two methods using the improved ACM with advanced texture features or using the conventional ACM in the multi-atlases framework, improve 0.4~2.3% and 5.8~9.9% for the 4 overlap metrics, compared to the method using only the conventional ACM.

4 Conclusion

In this paper, we have presented a learning-based method for accurate segmentation of hippocampus from the 7.0T MR images. Specifically, to handle the severe intensity inhomogeneity in the 7.0T images, the multi-atlases-based segmentation framework and the improved auto-context model (with advanced texture features) are employed for building multiple sequences of classifiers, and further used for segmentation of the new test images with label-fusion. The experimental results show that our model can achieve the significant improvement of the overlap scores, compared to a recent algorithm using only the auto-context model. In the future, we will extensively test our method using more 7.0T MR images that our collaborators will provide.

Reference

1. Zhou, J., Rajapakse, J.C.: Segmentation of subcortical brain structures using fuzzy templates. *NeuroImage* 28, 915–924 (2005)
2. Coupé, P., Manjón, J.V., Fonov, V., Pruessner, J., Robles, M., Collins, D.L.: Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *NeuroImage* 54, 940–954 (2011)
3. Khan, A.R., Wang, L., Beg, M.F.: FreeSurfer-initiated fully-automated subcortical brain segmentation in MRI using Large Deformation Diffeomorphic Metric Mapping. *NeuroImage* 41, 735–746 (2008)
4. Chupin, M., Hammers, A., Liu, R.S.N., Colliot, O., Burdett, J., Bardinet, E., Duncan, J.S., Garnero, L., Lemieux, L.: Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: Method and validation. *NeuroImage* 46, 749–761 (2009)
5. van der Lijn, F., den Heijer, T., Breteler, M.M.B., Niessen, W.J.: Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts. *NeuroImage* 43, 708–720 (2008)
6. Lötjönen, J.M.P., Wolz, R., Koikkalainen, J.R., Thurfjell, L., Waldemar, G., Soininen, H., Rueckert, D.: Fast and robust multi-atlas segmentation of brain magnetic resonance images. *NeuroImage* 49, 2352–2365
7. Powell, S., Magnotta, V.A., Johnson, H., Jammalamadaka, V.K., Pierson, R., Andreasen, N.C.: Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures. *NeuroImage* 39, 238–247 (2008)

8. Wang, H., Das, S.R., Suh, J.W., Altinay, M., Pluta, J., Craige, C., Avants, B., Yushkevich, P.A.: A learning-based wrapper method to correct systematic errors in automatic image segmentation: Consistently improved performance in hippocampus, cortex and brain segmentation. *NeuroImage* 55, 968–985
9. Tu, Z., Bai, X.: Auto-Context and Its Application to High-Level Vision Tasks and 3D Brain Image Segmentation. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 32, 1744–1757 (2010)
10. Freund, Y., Schapire, R.E.: A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences* 55, 119–139 (1997)
11. Cho, Z.-H., Han, J.-Y., Hwang, S.-I., Kim, D.-s., Kim, K.-N., Kim, N.-B., Kim, S.J., Chi, J.-G., Park, C.-W., Kim, Y.-B.: Quantitative analysis of the hippocampus using images obtained from 7.0 T MRI. *NeuroImage* 49, 2134–2140 (2010)
12. Morra, J.H., Tu, Z., Apostolova, L.G., Green, A.E., Toga, A.W., Thompson, P.M.: Automatic Subcortical Segmentation Using a Contextual Model. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) *MICCAI 2008, Part I. LNCS*, vol. 5241, pp. 194–201. Springer, Heidelberg (2008)

Texture Analysis by a PLS Based Method for Combined Feature Extraction and Selection

Joselene Marques^{1,2} and Erik Dam²

¹ University of Copenhagen, Denmark

² BiomedIQ, Copenhagen, Denmark

Abstract. We present a methodology that applies machine-learning techniques to guide partial least square regression (PLS) for feature extraction combined with feature selection. The developed methodology was evaluated in a framework that supports the diagnosis of knee osteoarthritis (OA). Initially, a set of texture features are extracted from the MRI scans. These features are used for segmenting the region-of-interest and as input to the PLS regression. Our method uses PLS output to rank the features and implements a learning step that iteratively selects the most important features and applies PLS to transform the new feature space. The selected bone texture features are used as input to a linear classifier trained to separate the subjects in healthy or OA. The developed algorithm selected 18% of the initial feature set and reached a generalization area-under-the-ROC of 0.93, which is higher than established markers known to relate to OA diagnosis.

Keywords: machine learning, PLS, classification, feature extraction, feature selection, texture analysis, OA, bone structure.

1 Introduction

Osteoarthritis (OA) is a widespread, complex disease that affects multiple components of the joint. The major OA characteristics are degradation of cartilage, changes in trabeculae bone structure, bone marrow lesions and osteophytes. The symptoms are swelling, pain, stiffness, and decreased mobility.

MRI can visualize the whole joint, allowing insights into OA diagnosis by detecting early tissue changes. Texture analysis applied to the MR images may capture these changes and provide the means for obtaining information that might not be assessed visually [1]. The comparison by Herlidou et al [2] suggests that diagnoses established by visual image analysis have major variability, whereas automated texture analysis provides a more accurate discrimination between control and pathologic subjects with skeletal muscle dystrophy.

Texture can be analyzed in terms of model-based signal processing, variation of intensity or structural elements. Kovaleva et al [3] applied extended multi-sort cooccurrence matrices to analyze structural brain asymmetry. The reported texture-based method was based on intensity, gradient, and anisotropy features. The investigation presented in [4], employed four automated methods of texture

analysis for structural characterization of trabecular bone. Another example is the study from Sørensen et al [5], where a general non-committed statistical machine-learning framework was used for measuring emphysema in CT images of the lungs.

These approaches combine different parameters and techniques to allow a broad representation of the image, but the drawback is the outcome of a potentially high-dimensional data set. Beyer showed that increasing the number of features leads to losing the meaning of each feature and possibly to decrease the model accuracy [6].

In many situations, a large number of features are correlated, not introducing new information to improve the ability to analyze the images. Feature selection and feature extraction are effective approaches to data reduction. Feature selection is the process that reduce the dimensionality by selecting a subset of original variables while feature extraction transforms and possibly reduces the dimension by some functional mapping of a D-dimensional vector onto a d-dimensional vector with ($d \leq D$) [7]. These techniques can provide a better understanding of the information being analyzed, besides decreasing the model complexity.

This study presents a new method that uses partial least squares (PLS) regression and machine-leaning to guide a combined feature selection and feature extraction algorithm. The PLS provides information to rank the features while the learning step iteratively transforms the feature space aiming for classification improvement and dimensionality reduction. A fully automatic framework applies this method to a texture feature bank extracted from low-field MRI scans of knees to classify subjects in healthy or with OA based on bone structure.

2 Background

2.1 Partial Least Squares in Classification Problems

PLS regression is a multivariate data analysis (MDA) technique that can be used to relate several response variables to several explanatory variables [8]. Concerning the classification problems, the response variables (y) are the group labels and explanatory variables (X) are the feature set.

The method aims to identify the underlying latent factors which best model the groups. Therefore, a linear composition of X is built to maximize the covariance between the X and y space, correlating the response variable with all the explanatory ones at once.

Our framework uses the SIMPLS algorithm for the PLS regression. Thus, for n samples, a single response y , and p predictors, a PLS regression model with k ($k < h$) latent variables can be expressed as follows:

$$X = TP' + E$$

$$y = UQ' + F$$

where $X \in \mathbb{R}^{n \times p}$ are the predictors, $T \in \mathbb{R}^{n \times h}$ and $U \in \mathbb{R}^{n \times h}$ are the X and y scores, $P \in \mathbb{R}^{p \times h}$ and $Q \in \mathbb{R}^{m \times h}$ are the X and Y loadings, $y \in \mathbb{R}^{n \times m}$ are the responses. The matrices $E \in \mathbb{R}^{n \times p}$ and $F \in \mathbb{R}^{n \times m}$ are residuals.

The latent factors (equation 1) and regression coefficients (equation 2) are computed based on the weight matrix $W \in \mathbb{R}^{pxk}$ that expresses the correlation of each X column with Y variable. Thereby, w -elements with values close to zero express less important explanatory variables:

$$T = XW \quad (1)$$

$$B = WQ' \quad (2)$$

2.2 Feature Selection Based on PLS

Several approaches for PLS based variable selection have been proposed. In a recent work, Li et al [9] proposed the competitive adaptive reweighted sampling (CARS) algorithm that uses the absolute values of regression coefficients of the PLS model as an index for evaluating the importance of wavelengths of multi-component spectral data.

Another example is the interactive variable selection (IVS) [10] that modifies the PLS algorithm by doing a dimension-wise selective reweighting of single values in the weight vector w . Their experiments showed that the elimination of either small or large values in w improved the model, but no clear explanation was given justifying the elimination of large w -values. Besides this, the author used a small simulated data set and the method was partly interactively evaluated.

Wold at all [11] introduced the variable importance in the projection (VIP), which is a score that summarizes the importance of each variable for the latent factors projections. The score is a weighted sum of squares of the weights in W , with the weights calculated from the amount of Y -variance of each PLS latent factor (see equation 3).

$$VIP_j = \sqrt{p \sum_{k=1}^h \left(SS(q_k t_k) (w_{jk} / \|w_k\|)^2 \right) / \sum_{k=1}^h (SS(q_k t_k))} \quad (3)$$

$$\text{where } SS(q_k t_k) = q_k^2 t_k' t_k$$

Building on Wold's work, Bryan et al [12] presented the MetaFIND application that implements a post-feature selection method based on VIP and correlation analysis of metabolomics data. The features were ranked, but the threshold that defines the selected features was a user-defined parameter.

Although PLS combined with VIP scores is often used when the multicollinearity is present among variables [10,11], there are few guidelines about how to use it [13]. The main contribution of this paper is a fully automatic PLS based strategy for dimensionality reduction (DR) combining feature extraction and feature selection.

Furthermore, the feature ranking function differs slightly from related work. Lindgren et al [10] suggests weighting the w -elements with their correlation to y . Similar to this; we use a ranking function that weights the w -elements with the percentage of variance explained by each latent factor. The next section describes our method.

3 The Dimensionality Reduction Method

The DR method uses cross validation and leave-one-out (LOO) to estimate the free parameters: the number of PLS latent factors and the number of features selected. The LOO on the training set provides data for calculating the PLS factors and evaluating the training classification, thereby defining the values of the free parameters to be used during the generalization.

3.1 Feature Selection and Extraction

As a pre-processing step, the feature set is auto-scaled before be submitted to the PLS algorithm. The outcome of the PLS regression depends on the scaling of data; therefore, this step gives the same importance to all features. A typical approach is to: (i) centre them by subtracting their averages, and (ii) scale each feature to unit variance by dividing them by their standard deviation.

In the next step, the PLS regression manipulates the training set considering all the features to model the structure of the feature space and their respective labels. From the outcome, the features are ranked using the index of equation 4:

$$\text{Index} = |W| * V \quad (4)$$

where $W \in \mathbb{R}^{(pxh)}$ is the weight matrix, $V \in \mathbb{R}^{(hx1)}$ is the percentage of variance explained by each latent factor on X and Y, considering h as the number of PLS latent factors with the best evaluation computed by LOO over the training set.

Finally, by forward regression, the algorithm sequentially adds the features to the feature set one at a time, re-compute the latent factors and evaluates the classification using different numbers of latent factors. The best evaluation defines the features and the number of latent factors to the generalization process.

For each sample of the generalization set, the latent factors are computed based on a re-computed matrix W considering the rest of the generalization set combined with the training set (LOO approach) and the selected feature set (see equation 1).

4 Framework

In this section we present the developed framework, introducing a brief description of the data collection, features generation, ROI definition, classification and evaluation.

4.1 Data Collection

The data set consists of MRI of both left and right knees from 159 test subjects in a community-based, non-treatment study. After exclusion of scans due to acquisition artefacts, 313 knee scans were included. The population characteristics were: age 56 ± 16 (mean and standard deviation), 47% female, and 19% with radiographic OA ($\text{KL} > 1$). The subjects signed informed consent forms. The

study was conducted in accordance with the Helsinki Declaration II and European Guidelines for Good Clinical Practice. The study protocol was approved by the local ethical committee. Figure 1 shows one example of a sagittal slice of a MRI Knee.

All scans were scored using the Kellgren & Lawrence score [14] determined from radiographs by an experienced radiologist. The score ranges from 0 to 4, where KL 0 indicates a healthy knee, KL 1 borderline OA, and KL 2–4 defines a knee with moderate to severe OA. When dividing in healthy/borderline and OA knees, the percentage of knees in each group is 81% and 19%.

4.2 Data Set Generation

Features. Following an uncommitted approach, 178 generic texture features were extracted from the images. A feature set that demonstrated to provide good results for many patterns is the N -jet [5,15,16]. We included the 3-jet, based on the Gaussian derivative kernels including up to the third order. Furthermore, to allow modelling of complex texture, non-linear combinations of the Gaussian derivative features were included. Specifically, these combinations were the structure tensor [17] and Hessian eigen-values and vectors; and gradient vector and magnitude.

Three different scales were included to capture anatomy and pathology of varying sizes. A small scale was included to encompass, for example, the trabecular structure of the bone and larger-scale features were needed to handle larger structures such as bone marrow lesions (BML). The scales were therefore chosen as 1, 2, and 4 mm. We also did experiments with scales 0.5, 2 and 8mm and the results were qualitatively identical. BMLs can be 1-2 centimetres in size — roughly corresponding to the support of a Gaussian filter at scale 4mm.

Feature Scores. The features are calculated in each voxel. However, to capture both the feature level and variation across a region-of-interest (ROI), we summarize each extracted feature in three possible scores: the mean, the standard deviation, and the Shannon entropy.

When extracting the features at three scales and calculating the three feature scores for each ROI, the total number of features was 534. This large, generic multi-scale feature bank contains linear and non-linear features including features that are invariant to rotation and intensity. Thereby, we hope to allow quantification of significant bone structures visible in the images.

4.3 ROI Definition

A voxel classification algorithm, designed for cartilage segmentation [15], was generalized to also segment the tibia bone. From the segmentation, we applied erosion to exclude the cortical bone. The remaining is the trabecular bone, which was defined as the ROI (figure 1).

4.4 Classification and Evaluation

The framework uses linear discriminant analysis (LDA) to classify the data and area-under-the-ROC (AUC) to determine the accuracy of the method. Since the data is unbalanced with respect to the number of knees for each class (healthy/OA), cost functions such as the classification accuracy is inappropriate to evaluate this dataset.

Another strategy to encompass the diversity of the data, particularly with limited samples, was to divide the data randomly into training and generalization sets 300 times, and hence 300 evaluations were performed.

For evaluating the generalization set, the training set and the generalization set are joined to be evaluated using LOO, but only the AUCs of the generalization set samples are used to compute the median AUC.

5 Experiments and Results

We evaluate the performance of our method by including the dimensionality reduction method in the framework. So, we apply the DR to the 534 features data set generated accordly with Section 4.2 and the outcome was propagated to the classification and evaluation process.

To investigate the influence of the feature selection process on the model accuracy, we also evaluate the performance of the model without the feature selection step. In this experiment, the PLS regression manipulated the training set considering all 534 features to generate the latent factors and sequentially the LOO approach was applied to define the number of PLS latent factors sent as input to the classifier. In the last experiment, we replaced our feature selection by the CARS algorithm [9] to compare the performances.



Fig. 1. Automatically segmented tibia bone in light gray and the region-of-interest (the trabecular bone) in gold

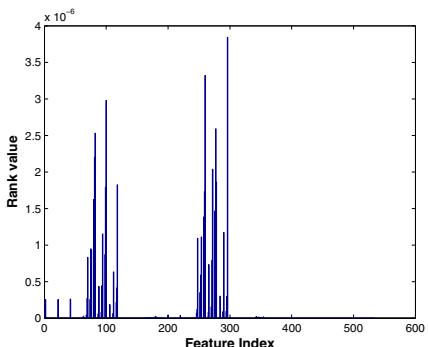


Fig. 2. Bar graph of the feature scores generated for one of the training sets. Based on the scores, the algorithm selected 95 of 534 features.

5.1 Results

In the first experiment, the ability to separate OA ($KL>1$) and healthy knees reached a generalization AUC of 0.93 ($p<0.0001$). The feature selection algorithm selected 93 features in median. Figure 2 shows the feature index generated for one of the training sets before being ordered.

Similarly, the second experiment, without the feature selection step, had generalization AUC of 0.91 ($p<0.0001$). The experiment using the CARS algorithm resulted in AUC of 0.86 with 43 features in median.

6 Discussion and Conclusion

The results showed that the presented texture analysis method captured trabecular bone changes and had diagnostic ability superior to other biomarkers of OA. The diagnosis using our PLS based method for dimensionality reduction reached AUC of 0.93. In a comparatively recent study using the same population [18], the Joint space width had AUC of 0.73 while the best individual marker, cartilage roughness, had AUC of 0.80. A linear combination of several morphometric and structural cartilage markers, cartilage longevity, scored AUC 0.84.

Our strategy for selecting features showed better than a recently developed method. The CARS algorithm selected less features, but decreased the accuracy of the model to AUC 0.86.

The experiment using feature extraction, but without the feature selection step had performance of AUC 0.91. This indicates that in spite of using less than 18% of the features, the developed feature selection step improved the model in terms of ability to analyze the images. Although the performance was only slightly better, the results suggest that our method detected some features that hinder the classification model instead of providing useful information.

Calculating all features used in the generic feature bank is time consuming. By including the feature selection step on the DR method we can identify the subset of features actually used by the framework and with a smaller subset, the overall processing time of the framework can be reduced.

Future improvements to the developed method include more pre-processing strategies to the PLS algorithm, for example, outlier treatment. Furthermore, we suggest validation on another populations or using high-field MRI scans. Besides this, the challenge that remains for future research is the examination on the relationship between the selected feature sets and the pathological features.

In conclusion, we present a dimensionality reduction method based on PLS regression that combines feature selection and feature extraction. The results illustrate that our method performs effectively to analyze texture and to reduce the number of features used by the framework.

Acknowledgments. We gratefully acknowledge the funding from the Danish Research Foundation (Den Danske Forskningsfond) supporting this work and the Center for Clinical and Basic Research for providing scans and radiographic readings.

References

1. Schad, L.R., Blüml, S., Zuna, I.: MR tissue characterization of intracranial tumors by means of texture analysis. *Magnetic Resonance Imaging* 11(6), 889–896 (1993)
2. Herlidou, S., Rolland, Y., Bansard, J.Y., Le Rumeur, E., de Certaines, J.D.: Comparison of automated and visual texture analysis in MRI: Characterization of normal and diseased skeletal muscle. *Magnetic Resonance Imaging* 17(9), 1393–1397 (1999)
3. Kovalev, V.A., Kruggel, F., von Cramon, D.: Gender and age effects in structural brain asymmetry as measured by MRI texture analysis. *NeuroImage* 19(3), 895–905 (2003)
4. Herlidou, S., Grebe, R., Grados, F., Leuyer, N., Fardellone, P., Meyer, M.E.: Influence of age and osteoporosis on calcaneus trabecular bone structure: a preliminary *in vivo* MRI study by quantitative texture analysis. *Magnetic Resonance Imaging* 22(2), 237–243 (2004)
5. Sørensen, L., Shaker, S.B., de Bruijne, M.: Quantitative analysis of pulmonary emphysema using local binary patterns. *IEEE Transactions on Medical Imaging* 29(2), 559–569 (2010)
6. Beyer, K.S., Goldstein, J., Ramakrishnan, R., Shaft, U.: When is "nearest neighbor" meaningful? In: Beeri, C., Bruneman, P. (eds.) *ICDT 1999. LNCS*, vol. 1540, pp. 217–235. Springer, Heidelberg (1998)
7. Liu, H., Motoda, H.: Feature Extraction, Construction and Selection: A Data Mining Perspective. Kluwer Academic Publishers, Norwell (1998)
8. Hubert, M., Branden, K.V.: Robust methods for partial least squares regression. *Journal of Chemometrics* 17(10), 537–549 (2003)
9. Li, H., Liang, Y., Xu, Q., Cao, D.: Key wavelengths screening using competitive adaptive reweighted sampling method for multivariate calibration. *Analytica Chimica Acta* 648(1), 77–84 (2009)
10. Lindgren, F., Geladi, P., Berglund, A., Sjostrom, M., Wold, S.: Interactive variable selection (IVS) for PLS. Part 1: Theory and algorithms. *Journal of Chemometrics* 8, 349–363 (1994)
11. Wold, S.: PLS-regression: a basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems* 58(2), 109–130 (2001)
12. Bryan, K., Brennan, L., Cunningham, P.: Metafind: A feature analysis tool for metabolomics data. *BMC Bioinformatics* 9 (2008)
13. Chong, I.G., Jun, C.H.: Performance of some variable selection methods when multicollinearity is present. *Chemometrics and Intelligent Laboratory System* 78(1–2), 103–112 (2005)
14. Kellgren, J.H., Lawrence, J.S.: Radiological assessment of osteo-arthrosis. *Annals of the Rheumatic Diseases* 16(4), 494–502 (1957)
15. Folkesson, J., Dam, E.B., Olsen, O.F., Pettersen, P.C., Christiansen, C.: Segmenting articular cartilage automatically using a voxel classification approach. *IEEE Transactions on Medical Imaging* 26, 106–115 (2007)
16. Florack, L., ter Haar Romeny, B., Viergever, M., Koenderink, J.: The Gaussian scale-space paradigm and the multiscale local jet. *International Journal of Computer Vision* 18(1), 61–75 (1996)
17. Weickert, J.: Anisotropic Diffusion in Image Processing. B.G.Teubner Stuttgart (1998)
18. Dam, E.B., Loog, M., Christiansen, C., Byrjalsen, I., Folkesson, J., Nielsen, M., Qazi, A., Pettersen, P.C., Garnero, P., Karsdal, M.A.: Identification of progressors in osteoarthritis by combining biochemical and MRI-based markers. *Arthritis Research & Therapy* 11(4), R115 (2009)

An Effective Supervised Framework for Retinal Blood Vessel Segmentation Using Local Standardisation and Bagging

Uyen T.V. Nguyen¹, Alauddin Bhuiyan¹, Kotagiri Ramamohanarao¹,
and Laurence A.F. Park²

¹ Department of Computer Science and Software Engineering,

The University of Melbourne, Australia

thivun@csse.unimelb.edu.au,

{abhuayan,kotagiri}@unimelb.edu.au

² School of Computing and Mathematics,

University of Western Sydney, Australia

lapark@scm.uws.edu.au

Abstract. In this paper, we present a supervised framework for extracting blood vessels from retinal images. The local standardisation of the green channel of the retinal image and the Gabor filter responses at four different scales are used as features for pixel classification. The Bayesian classifier is used with a bagging framework to classify each image pixel as vessel or background. A post processing method is also proposed to correct central reflex artifacts and improve the segmentation accuracy. On the public DRIVE database, our method achieves an accuracy of 0.9491 which is higher than any existing methods. More importantly, visual inspection on the segmentation results shows that our method gives two important improvements on the segmentation quality: vessels are well separated and central reflex are effectively removed. These are important factors that affect to the accuracy of all subsequent vascular analysis.

Keywords: Retinal images, blood vessel segmentation, central reflex, Gabor filter, Bayesian classifier, bagging.

1 Introduction

Recent research suggests that retinal vessel changes are important indicators for predicting cardiovascular diseases. For example, retinal arterial narrowing is directly related with hypertension and heart disease [1]. To quantify retinal vascular features for medical diagnosis, accurate segmentation of the vasculature from retinal images plays a critical role. Although a number of techniques have been proposed in the literature to address this task, ranging from the use of matched filter [2,3], tracking techniques [4], multi threshold probing [5], morphological operators [6], multi concavity analysis [7] or supervised learning [8,9,10] a significant improvement is still desirable. For example, visual inspection shows

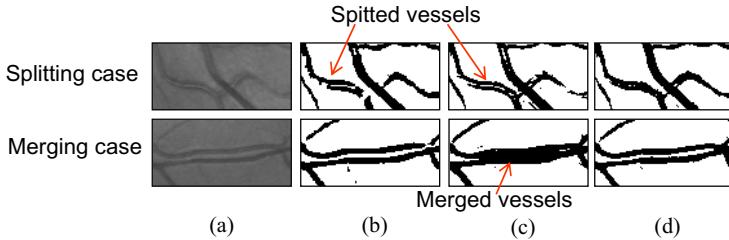


Fig. 1. Illustration of the limitations of existing methods: (a) two samples of a retinal image; segmentations of (b) Staal et al. method [8]; (c) Soares et al. method [9]; (d) the proposed method. The splitting of vessels due to central reflex are presented in Staal and Soares results while merging of close vessels happens in Soares result. The proposed method produces good results for both cases.

that segmentation results obtained by current approaches have two major limitations: 1) there is a high chance that close (or side-by-side) vessels are merged together and 2) the splitting of vessels due to the presence of central reflex (i.e. a bright strip along the center of the vessel). These two important factors, however, affect the accuracy of the subsequent analysis such as the identification of individual vessel segments, vessel mapping and vessel calibre measurement. The segmentation results from the state of the art are shown in Fig. 1 depicting these two limitations.

Motivated by this, we develop a supervised framework to improve retinal segmentation accuracy. The novelty of our approach lies in the use of local standardisation of the green channel, applying bagging to the classifier, and smoothing the central reflex. We show that each of these processes is beneficial to the accuracy of the segmentation.

2 Method

To perform supervised segmentation, we first obtain a set of features for each pixel, train the classifier using the pixel features and associated class value, and then apply any necessary smoothing of the results. In this section, we will begin by defining the features of each pixel, we will then proceed to describe the classifier, and then discuss the smoothing that was applied.

2.1 Pixel Features

For each pixel, five features are extracted including the local standardisation of the green channel and the maximum Gabor filter responses at four different scales. The details of each component are described as follows.

Locally standardised inverted green channel. The intensity of the green channel of a retinal image is used as a pixel feature due to its discrimination of vessel and non-vessel pixels. To overcome the illumination variation between

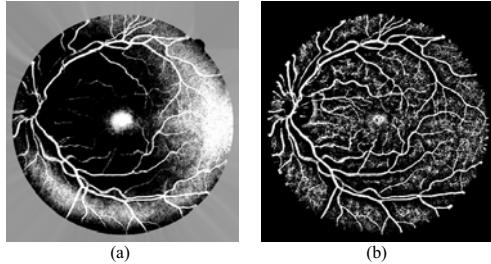


Fig. 2. (a) Global and (b) local standardisation applied to a retinal image

retinal images, the intensity levels of green channel are standardised to obtain a zero mean and unit standard deviation. We call this the global standardisation since the mean and standard deviation values are computed using the intensity levels of the whole image. The intensity levels of a retinal image after global standardisation is shown in Fig. 2a. An observation is that global standardisation does not help to overcome the non-uniform illumination in a retinal image (i.e. the vessel structures disappear in some regions and the background has a high intensity near the boundary of the fundus). So, instead of performing global standardisation, we standardize the image locally. This means that at each pixel position, a window of radius r centered on that pixel is identified. The standardization is then performed using pixels within this window to compute the value of the center pixel in the standardized image. Fig. 2b shows the local standardisation of the same image with the radius $r = 30$ pixels. This example shows that local standardisation provides greater discrimination between vessel and non-vessel pixels for this retinal image. Experiments were performed to identify the optimal radius and the results show that $r = 30$ gives the best separation between vessel and background classes. So, it is used as the first pixel feature in the proposed framework.

Gabor filter responses at different scales. The Gabor filter responses at different scales are used as additional pixel features due to its great ability of enhancing vessels while eliminating background noise [9]. The 2D Gabor filter is the modulation between a 2D Gaussian and a complex exponential [11]:

$$\psi(x, y) = \exp(j(k_0x + k_1y))\exp(-0.5(\frac{x^2}{\epsilon} + y^2)) \quad (1)$$

where $j = \sqrt{-1}$, k_0 and k_1 define the frequency of the complex exponential, ϵ defines the elongation of the Gaussian. The setting $k_0 = 0$, $k_1 = 3$ and $\epsilon = 4$ is used in our method as it has been shown experimentally by previous work [9] that this is an optimal setting for capturing vessel structure. Fig. 3 shows 2D images of the complex exponential, 2D Gaussian and the modulated 2D Gabor filter with this setting.

The Gabor filter response of an image is obtained by a convolution of the image with the Gabor filter. A filter bank consisting of Gabor filters with different

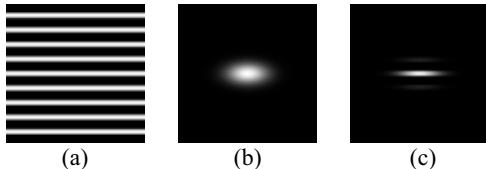


Fig. 3. 2D images of (a) the complex exponential; (b) 2D Gaussian; and (c) modulated Gabor filter with the setting $k_0 = 0$, $k_1 = 3$ and $\epsilon = 4$

orientations (θ) and scales (s) are used to identify vessels at different directions and widths. The Gabor filter response at a certain scale s is computed as the maximum response at that scale over all θ . Eighteen orientations (from 0° to 170° , with 10° of angular resolution) are considered to allow for vessels at different orientations. The final four features of each pixel are the the Gabor filter responses at four scales, s from 2 to 5.

2.2 Classification with Bagging and Bayesian Classifier

The novelty of the proposed classification method lies in the use of bagging [12] for improving the confidence in the classification prediction. From a set of training pixels, K classifiers are built using training pixels randomly sampled from the original training set. To classify an unlabeled pixel x , each classifier produces its own class prediction and the final class assigned to x is the majority of predictions returned by K classifiers. Different classifiers (i.e. k nearest neighbor, decision tree, random forest, support vector machines, and Bayesian classifier) were used as the base classifier but the results show that the Bayesian classifier with the class-conditional probability modeled by a mixture of Gaussians produces highest performance. Thus, it is used as the base classifier in our framework. Although bagging is a simple and popular technique for accuracy improvement in machine learning field, this is the first time it is introduced to retinal image segmentation area and has shown very promising results.

3 Smoothing Central Reflex Artifacts

After the classification phase, the posterior probability is thresholded to get the binary segmentation. However, for those images where the central reflex is evident (an example is shown in Fig. 4a), the pixels in the middle of the vessel are often misclassified as the background (as in Fig. 4b). So we propose here a smoothing technique for overcoming this problem. Our smoothing technique requires us to obtain an image (called complementary image) which is computed by a local differential operator performed on the inverted green channel of each image. At each pixel position (i, j) , the average gray levels of pixels along lines of length $L_1 = 3$ at four directions ($0^\circ, 45^\circ, 90^\circ, 135^\circ$) are evaluated and the maximum response is computed as $I_{max}(i, j)$. The average gray level of the pixels within a window of size $L_2 \times L_2$ ($L_2 = 15$) centered on the target pixel is also

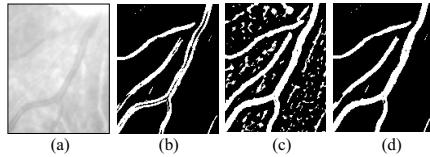


Fig. 4. Demonstration of smoothing to remove central reflex artefacts: (a) original image; (b) hard classification; (c) complementary image; (d) post-processed result

evaluated and defined as $I_{avg}(i, j)$. The response at pixel (i, j) in the complementary image is computed as $R(i, j) = I_{max}(i, j) - I_{avg}(i, j)$. The complementary image of Fig. 4a is shown in Fig. 4c.

An important property of the complementary image is that most of the central reflex pixels are correctly marked as vessel while the pixels between two nearby vessels are correctly labeled as background. However, this process also labels some background pixels as vessel due to the non-uniformity in the background of the retinal image. Removing this noise from the complementary image is a non-trivial task since they are connected to real vessels at some places in the image. The segmentations obtained by our classification method in Section 2 are, on the other hand, almost free of noise. By combining these two images, we can correct these misclassified pixels to overcome the central reflex problem. The combination process is as follows.

From the segmented image (obtained after the classification phase), all central reflex candidates (i.e. background pixels with high probability of being central reflex pixels) are identified. The central reflex probability of a pixel P_{ij} is computed as the fraction of the number of vessel pixels within its certain neighborhood. For example, if a neighborhood of size 3×3 is considered, this probability of pixel P_{ij} is identified as $\frac{1}{9}(\sum_{p=i-1}^{i+1} \sum_{q=j-1}^{j+1} N_{pq})$ where N_{pq} is a neighboring pixel of P_{ij} and $N_{pq} = 1$ if it is a vessel pixel and $N_{pq} = 0$ otherwise. The higher the number of vessel pixels around a background pixel, the higher probability of that pixel of being a central reflex pixel. The central reflex candidates are then identified as those with $P_{ij} > 0.5$. Each candidate pixel will be set as vessel in the final segmentation if it is also marked as vessel in the complementary image. This process is repeated iteratively until no changes are found in the final segmentation. Fig. 4d shows the final segmentation obtained after post processing procedure. Clearly, the method fills central reflex while keeping other vessel structures unchanged.

4 Experimental Results

The performance of our method is evaluated and compared to other methods on the publicly available DRIVE [8] databases. In this database, the segmentations of the first observer are used as the ground truth for evaluation while the performance of second observer is used as a benchmark for comparison. Training was performed on 20 training images and classification performance was measured

Table 1. Accuracy improvements with local standardisation, bagging and smoothing

Method	K1	K10	K1 & Smoothing	K10 & Smoothing
Global std.	0.9434	0.9466	0.9449	0.9475
Local std.	0.9464	0.9484	0.9473	0.9491



Fig. 5. An example demonstrating the contribution of bagging and post processing to the improvement in segmentation quality: (a) original image; segmentation obtained by our method (b) without bagging: ACC = 0.9388; (c) with bagging: ACC = 0.9443; (d) after smoothing: ACC = 0.9471

using 20 test images. To train a single classifier, 20000 pixels were randomly extracted from the training images. The bagging framework with 10 classifiers were used for all experiments. The accuracy (defined as the ratio of the number of correctly classified pixels to the total number of the pixels within the fundus area) was used as the main measure for evaluation and comparison.

4.1 Effect of Local Standardisation, Bagging and Smoothing

To understand the contribution of each proposed component to the performance of the whole system, a set of experiments was carried out to examine the accuracy of the method obtained without bagging (K1), with bagging (K10), and after smoothing. For each case, the accuracies obtained with local standardisation and global standardisation are reported. The results presented in Table 1 show that the average accuracies obtained by local standardisation are consistently higher than global standardisation while bagging consistently yields higher performance than the case of a single classifier. Smoothing process boosts the accuracy further. Paired Wilcoxon sign rank tests indicate that all of these differences are significant ($p < 0.0002$). These results mean that each proposed component is beneficial to the accuracy of the final segmentation result.

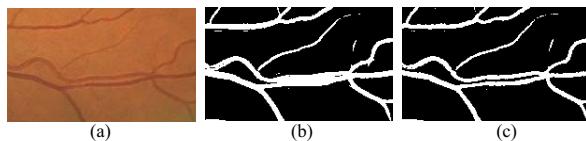
Fig. 5 visually demonstrates the effect of bagging and smoothing on improving the quality of a segmentation result. It is shown that bagging helps to smooth the vessel edge and remove false vessel detection while the smoothing component, as expected, helps to fill in vessels at those regions with central reflex.

4.2 Comparison to Existing Methods

The first column of Table 2 shows the average accuracy across 20 DRIVE images of different methods together with the accuracy of the proposed method before

Table 2. Average accuracy of different methods on DRIVE before and after smoothing and p-values of paired Wilcoxon sign rank tests

Method	ACC (before smoothing)	p-values (This work vs. other)	ACC (p-value) (after smoothing)
Second observer	0.9473	0.5256	0.9476 (0.00029)
Lam et al. [7]	0.9472	-	-
Mendonca et al. [6]	0.9463	-	-
Marin et al. [10]	0.9452	0.0003	-
Soares et al. [9]	0.9466	0.0008	0.9473 (0.00009)
Staal et al. [8]	0.9442	0.0002	0.9446 (0.00029)
Niemeijer et al. [13]	0.9416	0.0002	0.9420 (0.00019)
Proposed method	0.9484	-	0.9491 (0.00009)

**Fig. 6.** An example depicting the improvement in segmentation quality of proposed method: (a) original image; segmentation obtained (b) by Soares et al. method [9] ($ACC = 0.9509$) and (c) by our method ($ACC = 0.9646$). This improvement involves the correction of 334 pixels.

applying central reflex smoothing. The second column presents p-values of paired Wilcoxon sign rank tests obtained when comparing the proposed method with other methods (these values of some methods can not be computed due to the unavailability of their segmentation results). It is shown that our method gives highest average accuracy and it is comparable only with the second observer ($p = 0.5256$). The differences with all other methods are significant ($p < 0.001$).

Although the improvement in average accuracy is quite small, it should be highly appreciated for the fact that the accuracies obtained by current methods are already very high (which means that it is very hard to get any further improvement). More importantly, visual inspection on the segmentation results shows that these augmented true positives and true negatives help to improve the segmentation quality such as separating two nearby vessels and correcting central reflex pixels. These improvements need only a small increase in the number of pixels correctly classified (i.e. the separation of two nearby vessels in Fig. 6 involves the correction of only 334 pixels) but the quality improvement achieved is significant thanks to the correction in the vascular network that they provide.

We also applied our proposed smoothing technique to the segmentations produced by different methods to examine its ability on improving segmentation accuracy. The average accuracy together with p-value of significant test (obtained

when comparing individual accuracies achieved after and before smoothing) of each method are presented in the last column of Table 2. Obviously, the proposed smoothing method consistently boosts the accuracy of all methods and all improvements are significant ($p < 0.0003$). The average accuracy of the proposed method is also the highest among different methods, which demonstrates the effectiveness of the proposed framework on retinal vascular segmentation. In terms of running time, it takes about 36 minutes to train an ensemble of 10 Bayesian classifiers and less than 1 minute to segment a retinal image on a configuration of Duo CPU 2.39 GHz and 2Gb RAM.

5 Conclusions

We have proposed a supervised method for retinal blood vessel segmentation which is highly accurate and robust in the presence of vessel central reflex. Experiments show that the application of local intensity standardisation, bagging and central reflex smoothing have the strengths to 1) avoid merging of close vessels and 2) remove central reflex artefacts. These results have been observed in the segmentation results and an example is provided in Fig. 1. In the future, we will apply our segmentation method to map individual vessels and detect arteriovenous nicking to predict cardiovascular diseases.

References

1. Wong, T.Y., Klein, R., Sharrett, A.R., Duncan, B.B., Couper, D.J., Tielsch, J.M., Klein, B.E.K., Hubbard, L.D.: Retinal arteriolar narrowing and risk of coronary heart disease in men and women: the atherosclerosis risk in communities study. *Jama* 287(9), 1153 (2002)
2. Chaudhuri, S., Chatterjee, S., Katz, N., Nelson, M., Goldbaum, M.: Detection of blood vessels in retinal images using two-dimensional matched filters. *IEEE Transactions on Medical Imaging* 8(3), 263–269 (2002)
3. Hoover, A., Kouznetsova, V., Goldbaum, M.: Locating blood vessels in retinal images by piecewise threshold probing of a matched filter response. *IEEE Transactions on Medical Imaging* 19(3), 203–210 (2002)
4. Liu, I., Sun, Y.: Recursive tracking of vascular networks in angiograms based on the detection-deletion scheme. *IEEE Transactions on Medical Imaging* 12(2), 334–341 (2002)
5. Jiang, X., Mojon, D.: Adaptive local thresholding by verification-based multi-threshold probing with application to vessel detection in retinal images. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 25(1), 131–137 (2003)
6. Mendonca, A.M., Campilho, A.: Segmentation of retinal blood vessels by combining the detection of centerlines and morphological reconstruction. *IEEE Transactions on Medical Imaging* 25(9), 1200–1213 (2006)
7. Lam, B.S.Y., Gao, Y., Liew, A.W.C.: General retinal vessel segmentation using regularization-based multiconcavity modeling. *IEEE Transactions on Medical Imaging* 29(7), 1369–1381 (2010)
8. Staal, J., Abràmoff, M.D., Niemeijer, M., Viergever, M.A., van Ginneken, B.: Ridge-based vessel segmentation in color images of the retina. *IEEE Transactions on Medical Imaging* 23(4), 501–509 (2004)

9. Soares, J.V.B., Leandro, J.J.G., Cesar, R., Jelinek, H.F., Cree, M.J.: Retinal vessel segmentation using the 2-d gabor wavelet and supervised classification. *IEEE Transactions on Medical Imaging* 25(9), 1214–1222 (2006)
10. Marín, D., Aquino, A., Gegúndez, M., Bravo, J.: A new supervised method for blood vessel segmentation in retinal images by using gray-level and moment invariants-based features. *IEEE Transactions on Medical Imaging* (2010)
11. Movellan, J.R.: Tutorial on gabor filters. Open Source Document (2002)
12. Breiman, L.: Bagging predictors. *Machine Learning* 24(2), 123–140 (1996)
13. Niemeijer, M., Staal, J., van Ginneken, B., Loog, M., Abramoff, M.D.: Comparative study of retinal vessel segmentation methods on a new publicly available database. In: *Proceedings of SPIE*, vol. 5370, p. 648 (2004)

Automated Identification of Thoracolumbar Vertebrae Using Orthogonal Matching Pursuit

Tao Wu, Bing Jian, and Xiang Sean Zhou

Siemens Healthcare, SYNGO R&D USA

taowu@umiacs.umd.edu, {bing.jian,xiang.zhou}@siemens.com

Abstract. A reliable detection and definitive labeling of vertebrae can be difficult due to factors such as the limited imaging coverage and various vertebral anomalies. In this paper, we investigate the problem of identifying the last thoracic vertebra and the first lumbar vertebra in CT images, aiming to improve the accuracy of an automatic spine labeling system especially when the field of view is limited in the lower spine region. We present a dictionary-based classification method using a cascade of simultaneous orthogonal matching pursuit (SOMP) classifiers on 2D vertebral regions extracted from the maximum intensity projection (MIP) images. The performance of the proposed method in terms of accuracy and speed has been validated by experimental results on hundreds of CT images collected from various clinical sites.

1 Introduction

Accurate identification of the vertebral column structures is a crucial step in planning an image-guided spine surgery. The classic groupings of vertebral column structure include seven cervical (C1-C7), twelve thoracic(T1-T12), five lumbar (L1-L5), five sacral (S1-S5), and four coccygeal vertebrae. A typical practice of labeling vertebral column is either to count in a caudad direction from C2 or to count in a cephalad direction from L5, assuming the normal structure of the vertebral column. Due to the high prevalence of anomalies in the lumbosacral region, such as the sacralization of L5 and the lumbarization of S1 [1], the accurate detection of L5 and S1 is not trivial; therefore, counting from C2 inferiorly is usually considered more reliable. However, not all the imaging routines cover the C2 or even the cervicothoracic region especially for the diagnosis in the abdomen and/or pelvic region. In these scenarios the correct identification of thoracolumbar vertebrae, in particular, the last thoracic vertebra and the first lumbar vertebra, can be very helpful in making a definitive labeling. Because of the resemblance in the last thoracic vertebra and the first lumbar vertebra, and other factors such as the low spatial resolution and contrast in the imaging, an automated algorithm to assist differentiating the last thoracic and the first lumbar vertebra would be challenging but valuable.

Although there is a large body of research work on automating the analysis and interpretation of the spine image acquired on various imaging modalities, e.g., X-ray [2, 3], MR [4–9] and CT [9–13], most of these techniques focus

on the detection and segmentation of vertebral column structures; to our best knowledge, none of these methods was devoted to address the need and the development of accurately identifying thoracolumbar vertebrae which is the main contribution of this paper. The purpose of our study was to develop and test a computer algorithm for the identification of thoracolumbar vertebrae towards a seamless integration into an internal proprietary system for automatic vertebrae labeling. The labeling results are represented by a sequence of labeled vertebral bodies, each with the estimated center location and the orientation along the spinal curve geometry. For each labeled vertebra that lies in a specified thoracolumbar region (e.g. from T10 down to L2), the proposed classification algorithm checks if the input vertebra looks more like a thoracic vertebra or a lumbar vertebra. Furthermore, a voting scheme combines the classification results on these neighboring vertebrae can be used to consolidate the decision that if the input labeling is correct or not. We emphasize that the proposed vertebrae classification algorithm is compatible with any automatic labeling system providing the similar output interface regardless the internal implementation.

The rest of this paper is organized as follows: The technical description of the proposed method is elaborated in Section 2. Section 3 presents the experimental results on hundreds of CT volumes collected from various clinical sites. Finally, we draw conclusions in Section 4 with a brief discussion on the limitations and potential extensions of the proposed method.

2 Method

In this paper, we present a method to identify the thoracolumbar vertebrae using simultaneous orthogonal matching pursuit (SOMP) which is a recently developed greedy matching pursuit algorithm [14] and has shown excellent discriminative power in solving sparse approximation problems arising from many recognition applications. The proposed method is essentially a dictionary-based classification approach. At the training stage, dictionaries that capture the appearance characteristics of thoracic and lumbar vertebrae are trained independently from the human-annotated samples. At the testing stage, testing samples are identified by comparing the reconstruction errors after they are projected on the dictionary of each class. The details of different modules adopted in our approach, including the MIP generation, vertebra detection, dictionary building and the final classification, are described in the remaining part of this section.

2.1 MIP Generation and Vertebra Detection

Aiming to a highly efficient classification algorithm, we decide to investigate a 2D approach which can significantly reduce the computational time, compared to a fully 3D approach. By further considering the variability in the 2D axial views because of the projecting direction of ribs as shown in Fig. 1, we choose to look into the maximum intensity projection (MIP) images. The projection is performed in a volume of interest (VOI) containing the target vertebra body.

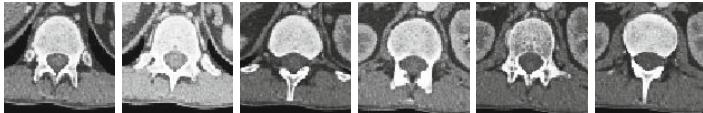


Fig. 1. The axial views intersecting the vertebral bodies of T12 (the left three) and L1 (the three on the right), respectively

The assumption is that, for each vertebra to be identified, we already have a good estimate of the body center location and the orientation along the spinal curve geometry, which can be used to estimate the boundary of the VOI as well as the projection direction for the MIP generation.

After the MIP generation, the next step is to extract the vertebra regions which are normalized to the same scale and provide the compact and representative features in the classification stage. The automatic extraction is achieved by a classic adaboost-based detection approach [15] in which a cascade of vertebra detectors is learned during an off-line training stage and then applied in the real-time detection. At each level of the cascade the classifier is tuned to achieve a high true positive rate ($>99.9\%$) on the training set. Meanwhile a moderate false alarm rate ($<0.1\%$) can be tolerated since the false positive detection results can be easily discarded by taking advantage of the prior that the target vertebra body region should be roughly positioned near the center of the MIP image in most cases since these centers used in the MIP generation are either marked by human or detected by a spine labeling algorithm. Examples of the extracted T12 and L1 vertebra regions are shown in Fig. 3. In case no vertebra is detected, a fixed-size region around the image center is cropped as a failure protection. Finally, the detected regions are resized to 32×32 matrices and then vectorized to provide input for the consequent training and testing stages.

2.2 SOMP Based Classifiers

The SOMP based classifier consists of a training stage and a testing stage. At the training stage, we build a dictionary, D_T , for the thoracic vertebrae and a dictionary, D_L , for the lumbar vertebrae. Each dictionary can be represented by a matrix. The column vectors, called atoms, are generated by detecting vertebral regions from MIP images in training set followed by proper resizing and vectorization. At the testing stage, given an image I and a dictionary D , the SOMP projection at each iteration finds the atom in D that is most strongly correlated with the residual of I which is initialized to I itself and then is updated by subtracting off the contribution from the selected atom. Let D_i denote the i -th atom (i -th column vector) in D . The procedure of projecting an image I onto a dictionary D and computing the residual R is summarized in Algorithm 1.

The identification result is obtained by comparing the residual magnitudes after projecting the testing sample onto all dictionaries. For example, if for a testing vertebra image I , the residual R_T after projecting it on the D_T is less in magnitude than the residual R_L from the D_L , then the testing sample is

identified as a thoracic vertebra, and vice versa. To further measure the confidence associated with the identification result, we define the identification score as $\max(R_T, R_L)/\min(R_T, R_L)$ and only trust the identification result if this ratio is greater than a pre-specified threshold. To improve the identification performance, we also built a cascade of SOMP classifiers using different dictionaries. One simple-to-implement cascading strategy is to check the score at each stage. If this score is lower than the threshold pre-specified for this stage, then the classifier in the next stage is triggered, and the identification score returned by the latter classifier will be used. Finally, the neighboring classification results can be combined together to vote the final identification.

Data: Testing image I , a dictionary D , the number of iterations K

Result: Residual R

```

 $R \leftarrow I$ 
for  $j \leftarrow 1$  to  $K$  do
     $\hat{i} \leftarrow \arg \min_i \| (R - D_i^T R D_i) \|$ 
     $R \leftarrow R - D_{\hat{i}}^T R D_{\hat{i}}$ 
end

```

Algorithm 1. SOMP projection of one testing image onto one dictionary

3 Experimental Results

We tested our algorithm on three datasets. Dataset I consists of 205 CT volumes with the thoracolumbar regions manually labeled. These volumes were selected from studies acquired at different institutions and have a good amount of variability in patient populations and imaging protocols. Dataset II contains 149 high-resolution colon scans of slice thickness between 1mm to 2mm. Dataset III contains 265 low-dose whole-body CT scans obtained in PET/CT studies and most images were acquired at 5mm slice thickness. For each case in these datasets, we first ran an automatic spine labeling algorithm and then applied the preprocessing step described in Section 2.1 to produce vertebra MIP images for classification. Note that dataset I is a subset of the training set that was used for training the automatic spine labeling algorithm. The thoracolumbar annotations from dataset I were also used in training the 2D vertebral region detector from MIP images as described in Section 2.1. The cascaded SOMP dictionaries were also built on the samples from dataset I. The first stage of the cascade was built using about half of samples selected from dataset I, then it was tested on all samples from dataset I. The threshold that determines if the identification decision can be made at the first stage or should be escalated to the second stage is determined empirically.

Helped by a radiologist, we manually examined the spine labeling and the vertebrae classification results obtained from dataset II and dataset III. Among the 149 T12 and 149 L1 vertebrae in dataset II given by the spine labeling

algorithm, the classification corrected the labeling mistakes on three studies, two with sacralization, and one with lumbarization, all leading to the wrong detection of L5. Meanwhile, the classification did not catch one mislabeling caused by the lumbarization and also misclassified L1 as T12 in two studies where the presence of short lumbar rib structures attached to L1 was found, which may cause the confusion. However, we also noticed that the three corrections all had the final scores greater than 1.15 and the three misclassifications were among the 11 vertebrae with final scores lower than 1.12. Therefore, the negative effects to the labeling system caused by the misclassification can be largely reduced by raising the final decision score threshold to a certain level without lowering the sensitivity on most cases.

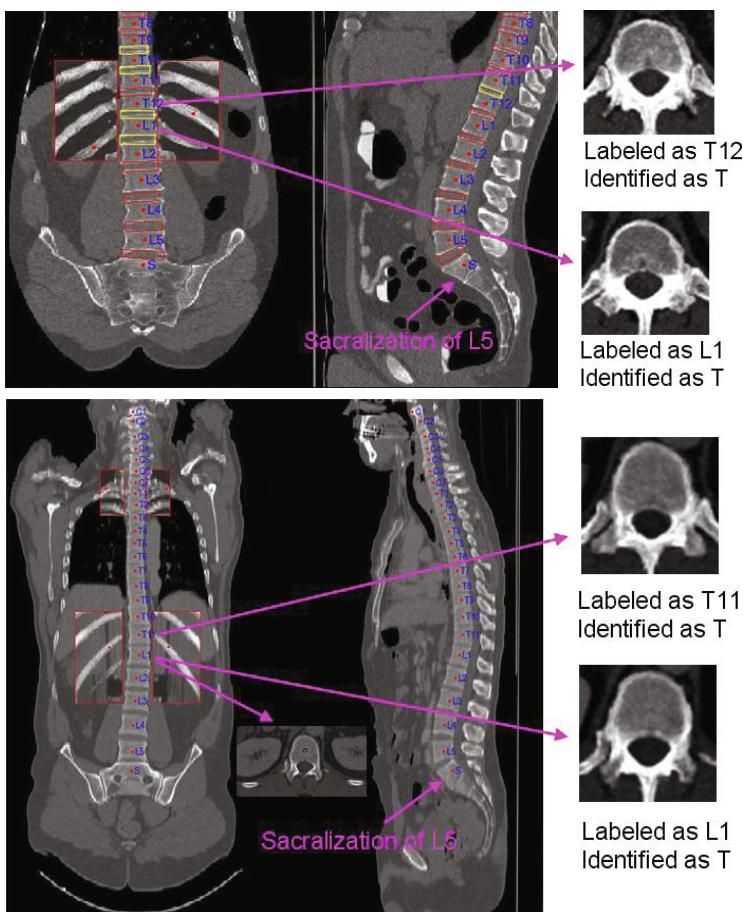


Fig. 2. Top: a colon scan in dataset II; Bottom: a whole-body scan in dataset III. In both cases, a sacralized L5 is found and leads to labeling error that can be corrected by vertebra classification results. Note that the rib attached to the true T12 on one side is hardly visible as shown in the embedded axial view.

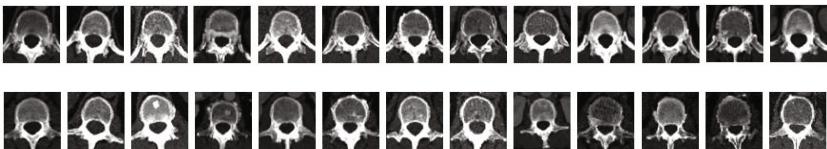


Fig. 3. Top row: examples from T12; Bottom row: examples from L1

We also checked how the vertebrae classification results agree with the spine labeling algorithm on dataset III. The conflicts on 29 vertebrae were found. The classification results were correct on 5 cases where the mislabelings due to the sacralization were confirmed. Among the 24 instances of misclassification, two had scores higher than 1.2 and both exhibit bilateral lumbar rib structures, three had scores between 1.1 and 1.2, and all others had scores below 1.1, including two cases with implants, one case with visible vertebral fracture and two cases with lumbar ribs on one side. We also noticed that among these 24 misclassifications there were 7 failures of the vertebral region detection due to the low image quality. Fig. 2 shows two studies where the labeling errors due to the sacralization of L5 were corrected by the vertebra classification module. Fig. 4 shows two studies where the vertebra classification module gave the correct results in the presence of lumbarization of S1. Fig. 5 shows several examples of misclassified vertebrae due to different reasons.

All experiments were performed on a PC with 2GB of RAM and a 2.0GHz Intel CPU. The average time for MIP generation and vertebral region detection is less than 0.2s per vertebra. The average time spent on one SOMP classifier is around 0.3s, and in our experiments, about 17% of the testing samples were sent to the second stage SOMP classifier, which may take additional 0.3s. Overall the average processing and classification time per vertebra is about 0.55s.

4 Conclusions

A fast and accurate method for recognition of the last thoracic vertebra and the first lumbar vertebra in CT scans is presented in this paper. The performance of the proposed method in terms of accuracy and speed has been validated on hundreds of CT images. Such a vertebra identification algorithm can be potentially integrated into an automated spine labeling system to improve the automation accuracy and hence provide better assistance in the radiology workflow. One major problem we found in experiments is the relatively high failure rate in handling the thoracolumbar transitional vertebrae, especially in the presence of various lumbar ribs¹, due to the current two-class setting. Since the SOMP method as a dictionary based algorithm is also suitable for multi-class identification problems, one of the future directions along this research line is to

¹ See <http://www.anatomyatlases.org/AnatomicVariants/SkeletalSystem/Images/15.shtml> for 20 varieties of lumbar ribs found on the L1

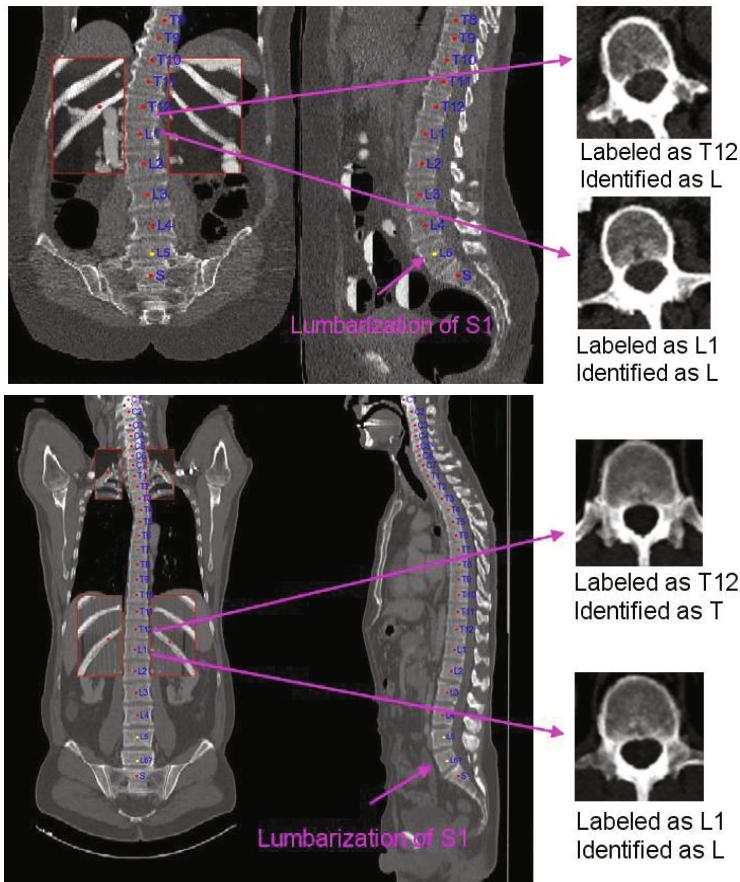


Fig. 4. Top: a colon CT scan from dataset II; Bottom: a whole-body CT scan from dataset III. Note that both cases exhibit the lumbarization of S1. The spine labeling algorithm wrongly located the L5 in the colon study due to the lumbarization, and the label propagation is found inconsistent with the vertebra classification results which are correct. In the whole-body scan, our spine labeling algorithm is able to detect that there might be six lumbar vertebrae and the labeling in the thoracalumbar region agrees with the vertebra classification results.

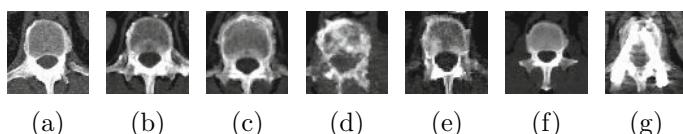


Fig. 5. Examples of misclassified vertebrae. (a-c) lumbar ribs; (d) vertebral fracture; (e) low image quality; (f) detection failure; (g) implant

extend the proposed method for detecting and identifying the different types of thoracolumbar and/or lumbosacral transitional vertebrae.

References

- [1] Konin, G.P., Walz, D.M.: Lumbosacral transitional vertebrae: Classification, imaging findings, and clinical relevance. *American Journal of Neuroradiology* 195, 465–466 (2010)
- [2] Smyth, P.P., Taylor, C.J., Adams, J.E.: Vertebral shape: automatic measurement with active shape models. *Radiology* 211, 571–578 (1999)
- [3] Dong, X., Zheng, G.: Automated vertebra identification from X-ray images. In: Campilho, A., Kamel, M. (eds.) *ICIPAR 2010 Part II*. LNCS, vol. 6112, pp. 1–9. Springer, Heidelberg (2010)
- [4] Peng, Z., Zhong, J., Wee, W.G., Lee, J.H.: Automated vertebra segmentation and quantification algorithm of the whole spine MR images. In: International Conference of the IEEE Engineering in Medicine and Biology Society (2005)
- [5] Schmidt, S., Kappes, J.H., Bergtholdt, M., Pekar, V., Dries, S.P.M., Bystrov, D., Schnörr, C.: Spine detection and labeling using a parts-based graphical model. In: Karssemeijer, N., Lelieveldt, B. (eds.) *IPMI 2007*. LNCS, vol. 4584, pp. 122–133. Springer, Heidelberg (2007)
- [6] Corso, J.J., Alomari, R.S., Chaudhary, V.: Lumbar disc localization and labeling with a probabilistic model on both pixel and object features. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) *MICCAI 2008, Part I*. LNCS, vol. 5241, pp. 202–210. Springer, Heidelberg (2008)
- [7] Huang, S.-H., Chu, Y.-H., Lai, S.-H., Novak, C.L.: Learning-based vertebra detection and iterative normalized-cut segmentation for spinal MRI. *IEEE Trans. Med. Imaging* 28(10), 1595–1605 (2009)
- [8] Kelm, B., Zhou, S., Suehling, M., Zheng, Y., Wels, M., Comaniciu, D.: Detection of 3D spinal geometry using iterated marginal space learning. In: Menze, B., Langs, G., Tu, Z., Criminisi, A. (eds.) *MICCAI 2010*. LNCS, vol. 6533, pp. 96–105. Springer, Heidelberg (2011)
- [9] Stern, D., Likar, B., Pernus, F., Vrtovec, T.: Automated detection of spinal centrelines, vertebral bodies and intervertebral discs in CT and MR images of lumbar spine. *Physics in Medicine and Biology* 55(1), 247 (2010)
- [10] Herring, J.L., Dawant, B.M.: Automatic lumbar vertebral identification using surface-based registration. *Journal of Biomedical Informatics* 34(2), 74–84 (2001)
- [11] Yao, J., O'Connor, S.D., Summers, R.M.: Automated spinal column extraction and partitioning. In: ISBI, pp. 390–393 (2006)
- [12] Klinder, T., Ostermann, J., Ehm, M., Franz, A., Kneser, R., Lorenz, C.: Automated model-based vertebra detection, identification, and segmentation in CT images. *Medical Image Analysis* 13(3), 471–482 (2009)
- [13] Ma, J., Lu, L., Zhan, Y., Zhou, X., Salganicoff, M., Krishnan, A.: Hierarchical segmentation and identification of thoracic vertebra using learning-based edge detection and coarse-to-fine deformable model. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) *MICCAI 2010 Part I*. LNCS, vol. 6361, pp. 19–27. Springer, Heidelberg (2010)
- [14] Tropp, J.A., Gilbert, A.C., Strauss, M.J.: Algorithms for simultaneous sparse approximation. Part I: Greedy pursuit. *Signal Processing* 86(3), 572–588 (2006)
- [15] Viola, P.A., Jones, M.J.: Rapid object detection using a boosted cascade of simple features. In: CVPR, vol. (1), pp. 511–518 (2001)

Segmentation of Skull Base Tumors from MRI Using a Hybrid Support Vector Machine-Based Method

Jiayin Zhou¹, Qi Tian¹, Vincent Chong², Wei Xiong¹, Weimin Huang¹,
and Zhimin Wang¹

¹ Institute for Infocomm Research, A*STAR, Singapore

² Department of Diagnostic Radiology, National University of Singapore, Singapore

Abstract. To achieve robust classification performance of support vector machine (SVM), it is essential to have balanced and representative samples for both positive and negative classes. A novel three-stage hybrid SVM (HSVM) is proposed and applied for the segmentation of skull base tumor. The main idea of the method is to construct an online hybrid support vector classifier (HSVC), which is a seamless and nature connection of one-class and binary SVMs, by a boosting tool. An initial tumor region was first pre-segmented by a one-class SVC (OSVC). Then the boosting tool was employed to automatically generate the negative (non-tumor) samples, according to certain criteria. Subsequently the pre-segmented initial tumor region and the non-tumor samples were used to train a binary SVC (BSVC). By the trained BSVC, the final tumor lesion was segmented out. This method was tested on 13 MR images data sets. Quantitative results suggested that the developed method achieved significantly higher segmentation accuracy than OSVC and BSVC.

1 Introduction

Tumor volume is not only a dominant prognostic indicator for many cancers, but also an objective measurement for the assessment of tumor's response to therapy. Measuring one-dimensional (1-D) [1] or two-dimensional (2-D) maximal diameters [2] in MR/CT images is clinically used to estimate tumor size. Previous studies have suggested however that real volume measurement provides more accurate estimates of the lesion sizes than 1-D and 2-D measurements. This observation gives more impacts for tumors with infiltrative and irregular pattern of growth, such as skull base, head and neck tumors. Thus tumor volumetry using robust and reliable segmentation methods is of great importance for prognosis, treatment planning and outcome evaluation.

For region-based segmentation methods, tumor segmentation problem can be considered as a region extraction problem. To discern tumor region from other non-tumor region, image voxels are classified/clustered into different tissue classes or groups according to certain similarity criteria. Typically this can be treated as a two-class problem to classify tumor and non-tumor classes. Support vector machine (SVM) is a supervised learning-based method and is primarily used for binary and also one-class and multi-class classifications. There were some studies for the segmentation of skull base and brain tumors from magnetic resonance (MR) images using either binary SVM (BSVM) classification [3,4,5] or one-class SVM (OSVM) recognition [6,7].

The common process in these works is to online learn the actual data distributions of target (tumor) and non-target (non-tumor) data by sampling, then train support vector classifiers (SVC) and extract the target data. To handle the non-linearly separable data classification problem, kernel mapping was used to transform the data into a higher dimensional feature space where a linear separation might be easier to achieve.

It is important to understand that the classification performance of either BSVC or OSVC is influenced by training samples as SVM is based on supervised learning. In order to learn the actual distribution properties of data explored, these “representative” training samples, which well reflect the distribution properties of the whole data, are needed. Indeed support vectors come from these “representative” training samples to construct either the optimal separating hyper-plane for binary classification or the enclosing hyper-sphere for one-class classification. On the other side, the selection of “representative” training samples may not be easy, especially the selection of negative samples for BSVC. For medical professionals who are the end-users of a supervised tumor segmentation system, it is probably not difficult to pick up tumor samples from images, however the manual selection of enough “representative” non-tumor samples is a tedious job because non-tumor data, which usually include highly diverse tissue types, occupy the majority portion in both image and feature spaces in most cases. In addition, the arbitrariness in the selection of non-tumor samples may cause considerable intra-/inter-operator variability in the final segmentation results. A common procedure is to adopt random selection of negative samples but the resultant classification performance is unstable. Some studies reported the use of OSVC to extract tumor region by learning data distribution from user-selected tumor samples only [6,7]. By this method, there is no need to select non-tumor samples. However, for heterogeneous tumors with blurry boundary, low true positive rates or high false positive rates were found in the final segmentation results. It is mainly due to the low discriminative power in OSVM-based data recognition.

Understanding the advantages/disadvantages of both OSVM and BSVM for image segmentation, it is interesting to design a complementary scheme which is able to utilize the advantages of both BSVM and OSVM and at the same time, to reduce their disadvantages. In this paper, we present a three-stage, hybrid SVM (HSVM)-based scheme for the segmentation of skull base tumor. In this method, H SVM is a seamless and natural connection of OSVM and BSVM with a boosting tool. The rest of the paper is organized as follows: In Section 2, we describe the details of the method to elaborate how HSVM works for tumor segmentation. In Section 3, the implementation of this method and the testing experiments are introduced. In Section 4, experimental results are reported and the performance of this method over other similar methods is demonstrated. The conclusion of this work is given in Section 5.

2 Method

2.1 Overview

The flowchart of the three-stage segmentation scheme is shown in Fig. 1. First an initial tumor region is pre-segmented by an OSVC. Then a boosting tool is employed to automatically generate the negative (non-tumor) samples, according to certain crit-

eria. The pre-segmented initial tumor region, which is treated as the positive samples, and the negative samples generated are used to train a BSVC, by which the final tumor lesion will be segmented out. The main idea of this scheme is to utilize the good discrimination capability of BSVM to be the main segmentation tool, while the good recognition capability of OSVM is utilized to be the guidance tool. Meanwhile, a boosting tool is used to connect the BSVM with the OSVM, by the automatic generation of negative samples.

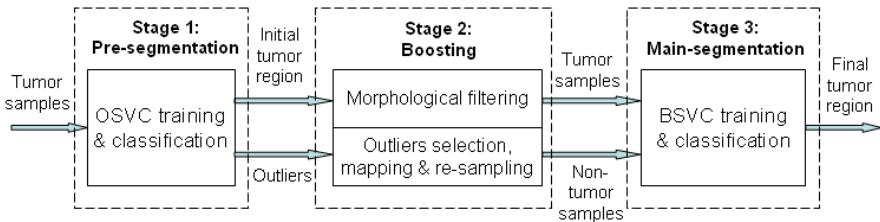


Fig. 1. The flowchart of the three-stage segmentation scheme

2.2 Pipeline of the HSVM Scheme

Stage 1 – Pre-segmentation: First an OSVC was trained by user-selected tumor samples. Then a rectangular ROI whose geometrical center was calculated by the coordinates of selected tumor samples was imposed to the same image. Thereafter the trained OSVC was used to extract the initial tumor region from this ROI, as shown in Fig. 2. For OSVC, the available training data are from only one of the classes, i.e., the target class (here is the tumor class) and there is no information about the other class, i.e., the outlier class, available. The task of OSVC is to define a boundary around the target class, such that it accepts as many of the targets as possible and excluding the outliers as many as possible. By an appropriate kernel mapping M that maps the data X to a higher dimensional feature space, a hypersphere can be sought to enclose the mapped target data $M(X)$ with a smallest radius R and center C , as shown in Fig. 3.

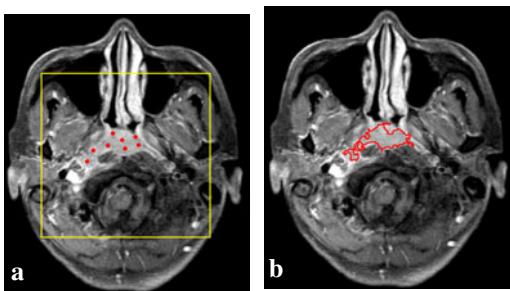


Fig. 2. a. The tumor samples selected and ROI;
b. initially segmented tumor region by OSVC

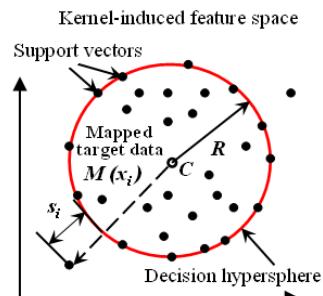


Fig. 3. Illustrator of OSVM with kernel mapping

Stage 2 – Boosting: In most of the cases, the extracted initial tumor region from Stage 1 is not a desired result. The majority of the extracted tumor region is inside the actual tumor region and its boundary has some distance to the actual tumor boundary. Hence another data classification procedure utilizing the discrimination capability of BSVC will follow up. The process of training a BSVC is equivalent to finding an optimal hyperplane in a way that minimizes the error on the training dataset and maximizes the perpendicular distance between the decision boundary and the closest data points in the two classes. The detailed mathematic descriptions of BSVC training, classification and implementation can be found in [8]. To well train a BSVC, both tumor and non-tumor samples are required. Here the initial tumor region extracted at Step 1 using OSVC, after morphological filtering, was used as the positive training samples. The negative training samples come from the “outliers” recognized by the OSVC at Stage 1. For the OSVC at Stage 1, data recognized as the target (tumor class) were enclosed by the optimally constructed hypersphere in the higher dimensional feature space whereas data not recognized as the target class scattered at the region outside the hypersphere. These outliers include non-tumor voxels, tumor voxels but unrecognized, and marginal voxels. The further an outlier is away from the hypersphere, the less similar it is with the tumor class and the less likely it belongs to the tumor class. Outliers which have high likelihood to be non-tumor voxels can be selected out for BSVC training, according to certain selection criteria. Specifically, suppose the radius of optimal hypersphere is R , those data which scattered at the region outside the concentric hypersphere (C, mR) with $m > 1$ were mapped backward into the image space, as shown in Fig. 4. They were used as the negative training samples after re-sampling, to equalize the numbers of samples from the positive class for better training of the BSVC in the next stage. In Fig. 4, solid spots inside the hypersphere (C, R) in feature space were mapped into image space as the positive samples (red region), and the hollow spots outside the hypersphere (C, mR) were mapped into image space and re-sampled as the negative samples (green scatters), which have a nearly uniform distribution in the non-tumor region.

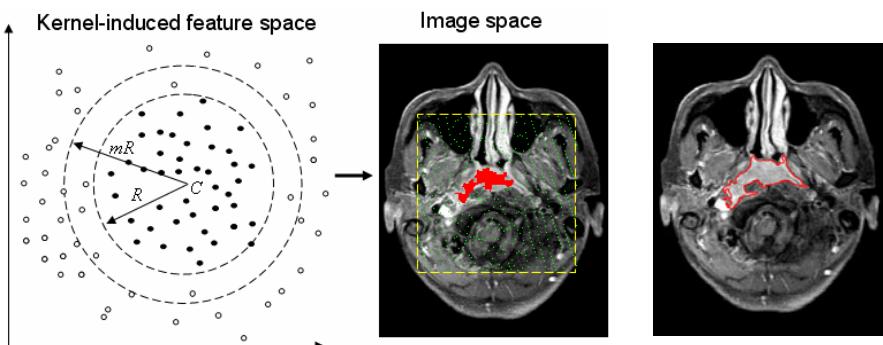


Fig. 4. The illustrator of mapping data spots from higher dimensional feature space to image spaces

Fig. 5. Tumor segmented in Stage 3

Stage 3 – Main-segmentation: A BSVC was trained using the positive and negative samples generated in Stage 2. Then the trained BSVC was applied into the ROI to segment the final tumor lesion by binary classification, as shown in Fig. 5. By this heuristic processing in Stage 2, more positives samples can be picked up for the training of the BSVC and the most important, an equal number of negative samples, which are very important for BSVC training, can be automatically generated. No user selection operation for negative samples is required.

2.3 Parameter Selection

Previous experience suggested that most of these real tumor voxels un-recognized by OSVC locate at the marginal area of the tumor mass which has the fuzzy transition to the marginal non-tumor area. Therefore before the backward-mapping, as shown in Fig. 4, parameter m ($m > 1$) was used to control the filtering of these possible marginal tumor voxels. Of course an excessively large m will filter out many true non-tumor voxels as well, hence the value of m needs to be tuned carefully such that the resultant new outliers include non-tumor voxels at a higher portion (ideal value, 100%) and tumor voxels at a lower portion (ideal value, 0). For this study, the proper value of m was determined by experiments which will be described in Section 3.

In addition, the Gaussian radius basis function (RBF) $K(\mathbf{x}, \mathbf{y}) = \exp(-\|\mathbf{x} - \mathbf{y}\|^2 / 2\sigma^2)$ was adopted as the learning kernel in this study. The kernel width parameter $2\sigma^2$ is used to reflect the degree of similarity of two data points. When increasing σ , the distance of the decision boundary to the support vectors increases and the number of support vectors decreases. Hence a proper σ gives a trade-off between the tight separating margin and the potential over-fitting. The online-learning scheme was adopted in this study hence for each slice to segment, the standard deviation calculated from the learning samples of tumor class was used as the σ . Another parameter is the rejection ratio $\nu \in (0, 1]$ in the OSVC at Stage 1. This user-specified parameter determines the ratio of points considered to be “outliers” in the target class. Similar to σ , ν also regulates the trade-off between the separating efficiency and the possible over-fitting problem. In this study, a ν value of 0.1 was used.

3 Experiment

Twenty-five MR data sets of patients with skull base tumors, including 20 cases of nasopharyngeal carcinoma (NPC) and 5 cases of hypophysoma, were used in this study. T1-weighted (T1W) and contrast-enhanced T1-weighted (CET1W, with fat suppression) images were acquire by a fast spin echo sequence. Images were acquired in axial or coronal plane and reconstructed to 512×512 pixels, field of view 200–230 mm; slice thickness of 5 mm without gap. The reference standards (RS) of tumors were manually traced out by an experienced head and neck radiologist.

In the implementation, signal intensities in T1W and CET1W images after median filtering within a 3×3 neighborhood were used as the input features. A greedy method was used to determine a proper value of m : For one pair of MR slices, tumor lesion was segmented by training an OSVC using operator selected tumor samples. Given

different m values, different sets of new outliers (non-tumor voxels) were obtained by the filtering of hypersphere (C, mR). Assume that in the ROI, voxel number of tumor RS is N_T , voxel number of non-tumor region is N_{N-T} , voxel number from new outliers but belonging to tumor RS is L_T , voxel number from new outliers but belonging to non-tumor region is L_{N-T} , and let $Co = L_{N-T}/N_{N-T} - L_T/N_T$. A higher value of Co means a more appropriate m , under the assumption that with a good m , the resultant outliers should hold a high portion of real non-tumor voxels and a low portion of real tumor voxels. Given individual m values from 1 to 1.6 with an interval of 0.05, an appropriate m value of 1.2 was determined by experiment using 81 pairs of tumor-containing slices with traced tumor RS from 12 MR data sets.

The remaining 13 MR data sets were used to test the proposed algorithm, with the benchmarking to OSVC and BSVC methods. Segmented tumors were compared by spatial voxel matching with RS . Two quantitative measures, volumetric overlap error (VOE, %) and average symmetric surface distance (ASSD, mm), were calculated to assess the similarity between the computerized and manually defined tumors [9]:

$$VOE = \left(1 - \frac{Vol_{Seg} \cap Vol_{RS}}{Vol_{Seg} \cup Vol_{RS}}\right) \times 100\% \quad (1)$$

$$ASSD = \frac{\sum_{a \in A} [\min_{b \in B} \{dist(a, b)\}] + \sum_{b \in B} [\min_{a \in A} \{dist(b, a)\}]}{N_A + N_B} \quad (2)$$

where A and B denote the surfaces of segmented and RS tumor volumes in each data set respectively, a and b are mesh points on A and B respectively, $dist(a, b)$ denotes the distance between a and b , N_A and N_B are the number of points on A and B . For VOE , a value is 0 is for a perfect segmentation and a value of 100% means that there is no overlap at all between segmentation and RS . $ASSD$ tells us how much on average the two surfaces differ and the value is 0 for a perfect segmentation. In addition, inter-operator variance (IV) was used to estimate the inter-operator reliability of each method at the voxel level:

$$IV = \left(1 - \frac{Vol_{Seg1} \cap Vol_{Seg2}}{Vol_{Seg1} \cup Vol_{Seg2}}\right) \times 100\% \quad (3)$$

where Vol_{Seg1} denotes segmented tumor volume by Operator 1, Vol_{Seg2} denotes the segmented tumor volume by Operator 2 using the same method. A lower IV value means the better inter-operator consistency.

4 Results

Totally 83 pairs of tumor-bearing MR slices from 13 MR data sets were tested in the experiment. Figures 6 and 7 show the images of one NPC lesion and one hypophysoma, including the original images with the manually traced tumor RS overlaid and the corresponding segmentation results using OSVC, BSVC and HSVC. In Fig. 6, NPC lesion invades not only parapharyngeal space but also prevertebral muscle and infratemporal fossa. The HSVC successfully extracted part of the lesion around the left infratemporal fossa while OSVC did not. All methods caught some

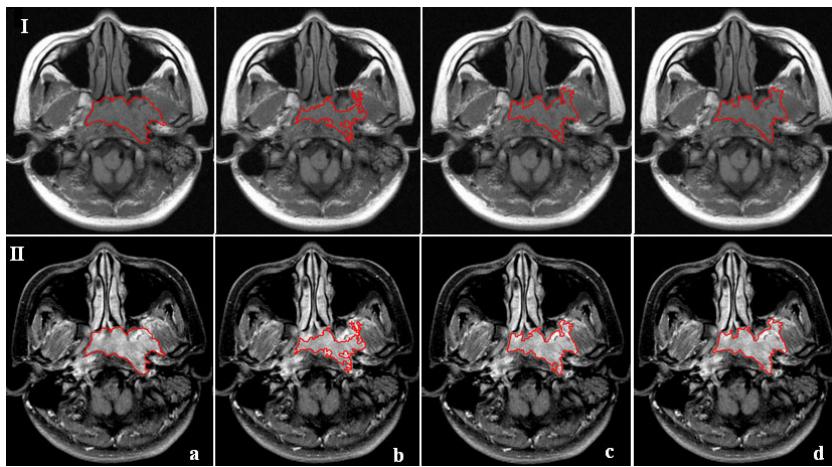


Fig. 6. Segmentation results for a NPC tumor. Row I: MR T1W image, Row II: CET1W image, Columns (a): tumor RS, (b)-(d): results of OSVC, BSVC and HSVC, respectively.

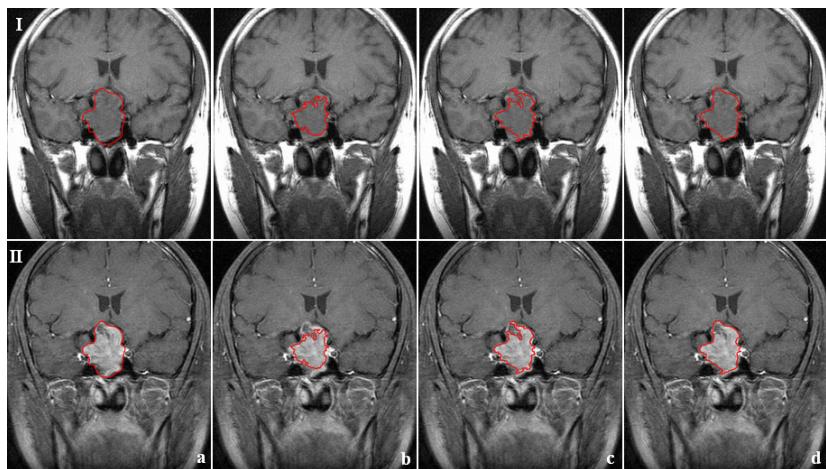


Fig. 7. Segmentation results for a hypophysoma. Row I: MR T1W image, Row II: CET1W image, Columns (a): tumor RS, (b)-(d): results of OSVC, BSVC and HSVC, respectively.

Table 1. Results of the evaluation metrics expressed as VOE, ASSD and IV

	VOE (%)			ASSD (mm)			IV (%)		
	OSVC	BSVC	HSVC	OSVC	BSVC	HSVC	OSVC	BSVC	HSVC
Min	22.8	21.4	15.0	0.9	0.9	0.3	8.5	13.7	12.6
Max	47.9	41.0	31.6	3.3	3.0	2.0	23.7	43.2	28.2
Mean	32.5	29.3	22.8	2.0	1.6	1.1	15.9	22.6	17.7
STD	7.6	5.6	5.2	0.7	0.6	0.5	4.3	6.9	4.6

false positives at the left lateral pterygoid muscles. From Fig. 7, it can be observed that compared to OSVC and BSVC, HSVC was able to identify the necrosis area (with lower signal intensity) which is generally included into the entire tumor volume. The quantitative validations at the voxel level for segmentation results using the three classifiers are summarized in Table 1. For HSVC, *VOE* and *ASSD* obtained were significantly lower than those obtained by using OSVC and BSVC ($p<0.05$, Kruskal-Wallis test). The OSVC obtained the best result of *IV* and the *IV* obtained by HSVC was slightly higher, however no significant difference was found. The *IVs* from OSVC and HSVC are significantly lower than that from BSVC ($p<0.05$, Kruskal-Wallis test). Both quantitative and visual results obtained showed that compared to OSVC and BSVC, the developed algorithm based on HSVC achieved better segmentation results.

5 Conclusions

A three-stage image segmentation method by exploring HSVC has been developed to conduct skull base tumor segmentation in MR images. In this method, a boost-ing tool was used to seamlessly connect an OSVC and a BSVC, by the automated generation and optimization of negative training samples. Implicitly the advantages of OSVC and BSVC were kept and some demerits were hidden, leading to the better classification performance for tumor region. Experimental results suggested that the developed method achieved better segmentation accuracy than OSVC and BSVC, and better inter-operator consistency than BSVC.

References

- Therasse, P., Arbuck, S.G., Eisenhauer, E.A., et al.: New guidelines to evaluate the response to treatment in solid tumors. *J. Natl. Cancer Inst.* 92, 205–216 (2000)
- Suzuki, C., Jacobsson, H., Hatschek, T., et al.: Radiologic measurement of tumor response to treatment: Practical approaches and limitations. *RadioGraphics* 28, 329–344 (2008)
- Zhou, J., Chan, K.L., Xu, P., Chong, V.F.: Nasopharyngeal carcinoma lesion segmentation from MR images by support vector machine. In: Proc. IEEE-ISBI, pp. 1364–1367 (2006)
- Ruan, S., Lebonvallet, S., Merabet, A., Constans, J.M.: Tumor segmentation from a multispectral MRI images by using support vector machine classification. In: Proc. IEEE-ISBI, pp. 1236–1239 (2007)
- Ayachi, R., Amor, N.B.: Brain tumor segmentation using support vector machines. In: Sossai, C., Chemello, G. (eds.) *ECSQARU 2009. LNCS*, vol. 5590, pp. 736–747. Springer, Heidelberg (2009)
- Zhang, J., Ma, K.K., Er, M.H., Chong, V.F.: Tumor segmentation from magnetic resonance imaging by learning via one-class support vector machine. In: Proc. IWAIT, pp. 207–211 (2004)
- Zhou, J., Chan, K.L., Chong, V.F., Krishnan, S.M.: Extraction of brain tumor from MR images using one-class support vector machine. In: Proc. IEEE-EMBC, pp. 6411–6414 (2005)
- Bishop, C.M.: *Pattern Recognition and Machine Learning*. Springer Science, New York (2006)
- Gerig, G., Jomier, M., Chakos, M.: Valmet: A new validation tool for assessing and improving 3D object segmentation. In: Niessen, W.J., Viergever, M.A. (eds.) *MICCAI 2001. LNCS*, vol. 2208, pp. 516–523. Springer, Heidelberg (2001)

Spatial Nonparametric Mixed-Effects Model with Spatial-Varying Coefficients for Analysis of Populations

Juan David Ospina^{1,2,4}, Oscar Acosta^{1,2}, Gaël Dréan^{1,2}, Guillaume Cazoulat^{1,2},
Antoine Simon^{1,2}, Juan Carlos Correa⁴,
Pascal Haigron^{1,2}, and Renaud de Crevoisier^{1,2,3}

¹ INSERM, U 642, Rennes, F-35000, France

² Université de Rennes 1, LTSI, F-35000, France

{Oscar.Acosta,Gael.Drean,Guillaume.Cazoulat,Antoine.Simon,
Pascal.Haigron}@univ-rennes1.fr

³ Département de Radiothérapie, Centre Eugène Marquis, Rennes, F-35000, France
r.de-crevoisier@rennes.fnclcc.fr

⁴ School of Statistics, Universidad Nacional de Colombia, Campus Medellín, Colombia
{jdospina,jccorrea}@unal.edu.co

Abstract. Voxel-wise comparisons have been largely used in the analysis of populations to identify biomarkers for pathologies, disease progression, or to predict a treatment outcome. On the basis of a good interindividual spatial alignment, 3D maps are produced, allowing to localise regions where significant differences between groups exist. However, these techniques have received some criticism as they rely on conditions which are not always met. Firstly, the results may be affected by misregistrations; secondly, the statistics behind the models assumes normally distributed data; finally, because of the size of the images, some strategies must be used to control for the rate of false detection. In this paper, we propose a spatial (3D) nonparametric mixed-effects model for population analysis. It overcomes some of the issues of classical voxel-based approaches, namely robustness to false positive rates, misregistrations and large variances between groups. Examples on numerical phantoms and real clinical data illustrate the feasibility of the approach. An example of application within the development of voxel-wise predictive models of rectal toxicity in prostate cancer radiotherapy is presented. Results demonstrate an improved sensitivity and reliability for group analysis compared with standard voxel-wise methods and open the way for potential applications in computational anatomy.

1 Introduction

Statistical voxel-based methods provide a way to reveal regional changes between groups by locally computing the difference of a signal across a given population [1]. Among these techniques, voxel based morphometry (VBM) [2] in its standard form, for example, assesses the differences between groups by fitting the general linear model (GLM) to the available signal from all subjects at each voxel and relates it to different covariates such as age, gender, diagnosis, cognitive scores, etc. Thus, the findings may be related with density changes of a given tissue such as the gray matter (GM) in brain

studies, for example [3]. These differences may also represent anatomical changes, encoded in Jacobian maps, such as in Tensor Based Morphometry (TBM) [4,5], obtained by warping individuals to a common template and assessing the atrophy or expansion after the non-rigid deformation. Some studies have combined hypometabolism and atrophy using PET, where the differences may be related with functional integrity [6,7] or combined with fMRI where they encode a response caused by an external stimulus [8]. More recent studies address the problem of producing spatial patterns after a treatment, aimed at predicting failure rates. This is the case of external beam radiotherapy for prostate cancer, where the relationship between the spatial pattern of the planned dose in radiotherapy may be correlated with toxicity events. The importance of spatial location in this clinical application has been already pointed out by [9]. More recently published papers already performed voxel wise analysis to statistically compare toxicity outcomes in the urinary tract [10] and tumour control in the prostate [11].

Voxel-based techniques have received some criticism in the past [12] and no agreement has been reached around some concerns [13]. In order to compare similar spatial patterns, one of the most challenging requirements of this methodology is the alignment of the whole population in the same anatomical space. Another issue deals with the generation and testing of statistics at a voxel level and the third challenge is the correction of these statistics accounting for false positives and the multiple comparison problem. To cope with misregistration issues, most of the approaches smooth the data thereby increasing the overlap across the individuals. However, this may reduce the sensitivity to detect and leads to a less precise localisation of regional differences as it ignores the spatial correlation structure. Concerning with the voxel-wise hypothesis testing, it may not fulfill the normality assumptions leading to a misspecification of the testing statistics and its distributions under the null hypothesis. On the other hand, the problem of multiple comparisons and of the false positive rate has been tackled by using Bonferroni correction, False Discovery Rate [14], Permutation testing [15] or Gaussian Random Fields [16] but sometimes at the expense of the number of false negatives.

In this paper, we propose a 3D nonparametric mixed-effects model for population analysis. As opposed to classical voxel-based approaches it exploits intra-individual spatial correlation at each voxel location leading to a separation of: i) the underlying average population behaviour; ii) the individual specificities; iii) the effect of any given variable as it will be shown below. The method proposed here builds upon the work of nonparametric mixed-effect models as introduced by [17]. Following the scheme proposed by [18] for longitudinal data, we extended the model to a spatial framework to detect specific patterns characterizing differences amongst several groups within the same population.

Nonparametric models are a broad class of techniques concerned with data modeling (probability distributions, regression models, etc.) by using the data itself to produce the data description with no intermediation of a functional descriptor of the relationship between input and output variables. Applications in image processing can be found for example in [19]. On the other hand, mixed-effects models have been applied in medical imaging for group analysis, for example in fMRI [20]. To our knowledge there have been no previous attempts to use the nonparametric mixed-effects model in medical

imaging for spatial population analysis, where each individual's data corresponds to a \mathbb{R}^3 -to- \mathbb{R} function.

This paper is structured as follows. After a brief introduction of the theoretical background, experiments with synthetic phantoms and real data are presented. An example of application within the development of voxel-wise predictive models of rectal toxicity in prostate cancer radiotherapy is proposed. Finally, the results obtained and the conclusions illustrate the benefits of the method and open the way to a broad range of clinical applications.

1.1 Statistical Model: Nonparametric Approach

The problem of two-group comparison can be seen as follows. Let's consider a collection of n independent individuals, divided into two different groups (T and its complement T^C), whose images were acquired or processed following the same protocol. The images are supposed spatially normalised to a single coordinate system and therefore resampled into the same lattice $D \subseteq \mathbb{R}^3$ containing d points. The goal is therefore to identify the differences between the two groups, at each point $\mathbf{x} \in D$. Let's assume that the response (image intensity), for the i -th individual at the point \mathbf{x} is $f_i(\mathbf{x})$, as in eq. (1),

$$f_i(\mathbf{x}) = m(\mathbf{x}) + g(\mathbf{x}) 1_{\{i \in T\}} + e_i(\mathbf{x}) + \varepsilon_i(\mathbf{x}), \quad (1)$$

where $m(\mathbf{x})$, $g(\mathbf{x})$ and $e_i(\mathbf{x})$ are smooth functions and $\varepsilon_i(\mathbf{x})$ is a random error. $1_{\{i \in T\}}$ is one if the i -th individual belongs to group T and it is zero otherwise. The term $m(\mathbf{x}) + g(\mathbf{x})$ represents the mean of the group T , and $m(\mathbf{x})$ the mean of the group T^C . Further, $e_i(\mathbf{x})$ is the deviation of the i -th individual from the corresponding mean of its group. Because this kind of model is composed of population and individual features, it has been referred to in the litterature as a *mixed-effects model*. It has to be noted that, although for a single individual i , conditioned with respect to himself, the function $e_i(\cdot)$ is deterministic; from the whole population perspective, this is quite a random feature.

The function $g(\cdot)$ could be thought of as the coefficient of the variable $1_{\{i \in T\}}$. Because $g(\cdot)$ may change throughout D , it is called a *spatial-varying coefficient*, representing the variation in the mean for the individuals belonging to the group T , with respect to the individuals belonging to the group T^C . As a result, the points where the function $g(\cdot)$ is not negligible, represent the areas where both groups are different.

Although the error term $\varepsilon_i(\cdot)$ is not assumed to be of any particular type of process, the correlation structure may be thought of as being spatial autoregressive as in eq. (2),

$$\text{Cov}(\varepsilon_i(\mathbf{x}), \varepsilon_j(\mathbf{y})) = \begin{cases} 0, & i \neq j \\ \sigma^2 \rho^{\|\mathbf{x}-\mathbf{y}\|}, & i = j \end{cases}, \quad (2)$$

where σ^2 is the variance of $\varepsilon_i(\cdot)$ and $\rho^{\|\mathbf{x}-\mathbf{y}\|}$ is the correlation between $\varepsilon_i(\mathbf{x})$ and $\varepsilon_i(\mathbf{y})$ which depends only on the distance between \mathbf{x} and \mathbf{y} under the norm $\|\cdot\|$. This correlation structure is adequate when for closer points, the measures of one individual are more correlated than for separate ones. As individuals are independent, the correlation between them is always zero.

1.2 Model Estimation

To estimate the value of $m(\mathbf{x})$, $g(\mathbf{x})$ and $e_i(\mathbf{x})$, at each point \mathbf{x} and the parameters σ^2 and ρ , the Naive Local Polynomial Kernel Method proposed in [17] was used. It was assumed that the first order approximation is valid for all of them around any point \mathbf{x}_0 , thus $m(\mathbf{x}) \approx m(\mathbf{x}_0) + \nabla m^T(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$, $g(\mathbf{x}) \approx g(\mathbf{x}_0) + \nabla g^T(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$ and $e_i(\mathbf{x}) \approx e_i(\mathbf{x}_0) + \nabla e_i^T(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)$. Then, the response $f_i(\mathbf{x})$ can be approximated as in eq. (3),

$$f_i(\mathbf{x}) \approx m(\mathbf{x}_0) + \nabla m^T(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + (g(\mathbf{x}_0) + \nabla g^T(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0)) 1_{\{i \in T\}} + e_i(\mathbf{x}_0) + \nabla e_i^T(\mathbf{x}_0)(\mathbf{x} - \mathbf{x}_0) + \varepsilon_i(x). \quad (3)$$

Suppose now that for the i -th patient the response variable is measured at points $\mathbf{x}_{t_1}, \mathbf{x}_{t_2}, \dots, \mathbf{x}_{t_d}$, for $i = 1, 2, \dots, n$. The index t_j , $j = 1, \dots, d$ is used to identify the points within a 3D-lattice with the components of a vector of length d . Thus, $n \times d$ points are available to estimate the model. The strategy to achieve the representation of the response in terms of the previously mentioned components is to estimate the model of eq. (3) at each point \mathbf{x}_{t_j} , using the expansion around the same point \mathbf{x}_{t_j} . Define $\boldsymbol{\beta} = [m(\mathbf{x}_{t_j}) \nabla m^T(\mathbf{x}_{t_j})]^T$, $\boldsymbol{\beta}^g = [g(\mathbf{x}_{t_j}) \nabla g^T(\mathbf{x}_{t_j})]^T$, $\mathbf{b}_i = [e_i(\mathbf{x}_{t_j}) \nabla e_i^T(\mathbf{x}_{t_j})]^T$ and $\Delta_{jk} = [1 (\mathbf{x}_{t_k} - \mathbf{x}_{t_j})^T]^T$. Then, if \mathbf{x}_{t_k} is close enough from \mathbf{x}_{t_j} , using eq. (3) it is possible to express the response variable for the i -th individual at the point \mathbf{x}_{t_k} as in eq. (4),

$$f_i(\mathbf{x}_{t_k}) = \Delta_{jk}^T \boldsymbol{\beta} + 1_{\{i \in T\}} \Delta_{jk}^T \boldsymbol{\beta}^g + \Delta_{jk}^T \mathbf{b}_i + \varepsilon_i(\mathbf{x}_{t_k}). \quad (4)$$

For each individual it is defined the vector $\mathbf{Y}_i = [f_i(\mathbf{x}_{t_1}) \dots f_i(\mathbf{x}_{t_d})]^T$, that in matrix notation can be expressed as in eq. (5),

$$\mathbf{Y}_i = \mathbf{X}\boldsymbol{\beta} + 1_{\{i \in T\}} \mathbf{X}\boldsymbol{\beta}^g + \mathbf{X}\mathbf{b}_i + \varepsilon_i, \quad (5)$$

$$\text{where } \varepsilon_i = [\varepsilon_i(\mathbf{x}_{t_1}) \dots \varepsilon_i(\mathbf{x}_{t_d})]^T \text{ and } \mathbf{X} = \begin{bmatrix} \Delta_{j1}^T \\ \vdots \\ \Delta_{jd}^T \end{bmatrix}.$$

Moreover, defining $\mathbf{Y} = [\mathbf{Y}_1^T \dots \mathbf{Y}_n^T]^T$, $\varepsilon = [\varepsilon_1 \dots \varepsilon_n]^T$, $\mathbf{b} = [\mathbf{b}_1^T \dots \mathbf{b}_n^T]^T$, $\mathbf{X}^* = \begin{bmatrix} \mathbf{X} \\ \vdots \\ \mathbf{X} \end{bmatrix}$, $\mathbf{Z} = \begin{bmatrix} \mathbf{X} & & \\ & \ddots & \\ & & \mathbf{X} \end{bmatrix}$ and $\mathbf{V} = \begin{bmatrix} 1_{\{1 \in T\}} \mathbf{X} \\ \vdots \\ 1_{\{n \in T\}} \mathbf{X} \end{bmatrix}$

it is possible to write the model for the whole data around \mathbf{x}_{t_j} , as in eq. (6),

$$\mathbf{Y} = \mathbf{X}^* \boldsymbol{\beta} + \mathbf{V} \boldsymbol{\beta}^g + \mathbf{Z} \mathbf{b} + \varepsilon. \quad (6)$$

The first component of the estimated vector $\boldsymbol{\beta}$ is the estimate of $m(\mathbf{x}_{t_j})$ and the first entry of the estimate of $\boldsymbol{\beta}^g$ is the estimate of $g(\mathbf{x}_{t_j})$. The remaining components correspond to the estimates of the gradient of both functions. The estimate of the vector \mathbf{b} contains the estimates of each $e_i(\mathbf{x}_{t_j})$ and its gradient. Despite the smoothness of

$m(\cdot)$, $g(\cdot)$ and $e_i(\cdot)$, $i = 1, \dots, n$, it is possible that for not so close points to \mathbf{x}_{t_j} , these functions behave differently. Thus, a kernel is used to weigh the contribution of each point \mathbf{x}_{t_k} , $k = 1, \dots, d$, when approaching the terms at \mathbf{x}_{t_j} . If the maximum local likelihood method is used (see [21]), the problem of estimating the parameters in eq. (6) can be reduced to estimating the parameters of the weighted model presented in eq. (7),

$$\mathbf{K}_h^{1/2} \mathbf{Y} = \mathbf{K}_h^{1/2} \mathbf{X}^* \boldsymbol{\beta} + \mathbf{K}_h^{1/2} \mathbf{V} \boldsymbol{\beta}^g + \mathbf{K}_h^{1/2} \mathbf{Z} \mathbf{b} + \boldsymbol{\varepsilon}', \quad (7)$$

where $\mathbf{K}_h = \text{diag}(\mathbf{K}_{jh}, \dots, \mathbf{K}_{jh})$ is a $nd \times nd$ matrix and

$$\mathbf{K}_{jh} = \text{diag}(k_h(\|\mathbf{x}_{t_1} - \mathbf{x}_{t_j}\|), \dots, k_h(\|\mathbf{x}_{t_N} - \mathbf{x}_{t_j}\|))$$

is a $d \times d$ matrix whose entries are obtained using $k_h(\cdot) = \frac{1}{h} k\left(\frac{\cdot}{h}\right)$, a kernel function with bandwidth h . This is usually found in the literature as a local mixed-effects model [18]. It is a well-known fact that no matter the selected kernel function, the critical step in the nonparametric procedures is the bandwidth selection [18]. In this last reference, the authors present various criteria based on *leave-one-out* and *cross-validation*, that can be immediately applied in this context.

The presented method can be viewed as a generalization of previous works on voxel-based differences testing. For example, using a zero order approximation for $f_i(\cdot)$ as well as assuming that $e_i(\cdot)$ is always zero for all the individuals (meaning that there are no individual effects) and choosing conveniently the kernel function, it can be shown that $g(\mathbf{x}_{t_k})$ is the difference, at \mathbf{x}_{t_k} , in the means of individuals belonging to the T group and of those that do not. In addition, dividing $\hat{\beta}_1^g$ (the estimate of $g(\mathbf{x}_{t_k})$) by its standard deviation (say $sd[\hat{\beta}_1^g]$, obtained via the fitting procedure under these hypothesis) the t-statistics emerges. Furthermore, if these hypothesis remain except for a Gaussian kernel being used and an arbitrary value allowed for h , then $\hat{\beta}_1^g / sd[\hat{\beta}_1^g]$ is the same statistic used in [16].

2 Experiments and Results

Two different scenarios were devised to evaluate the performance and potential applications of the proposed method. For the first one, we used numerical phantoms and for the second one, real data coming from a study of rectal bleeding after prostate cancer treatment with external beam radiotherapy. In both cases we used the Epanechnikov kernel, with a norm function $\|\mathbf{x} - \mathbf{y}\| = \max_i |x_i - y_i|$ and h accordingly selected to minimize the functional $\sum_{ij} \left(\hat{f}_i(\mathbf{x}_{t_j}) - f_i(\mathbf{x}_{t_j}) \right)^2$, where the $\hat{f}_i(\mathbf{x}_{t_j})$ denotes the predicted

value for the i -th individual at \mathbf{x}_{t_j} and $f_i(\mathbf{x}_{t_j})$ the actual value. The methodology was implemented using R software [22] and the function *lme()* from the *nlme* package [23] for running the local mixed-effects regressions. The statistical significance was evaluated using standard theory (Wald test), only for the results concerning the estimation of $g(\mathbf{x})$.

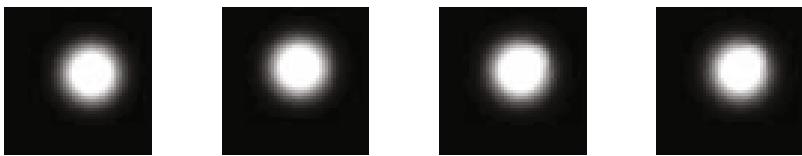
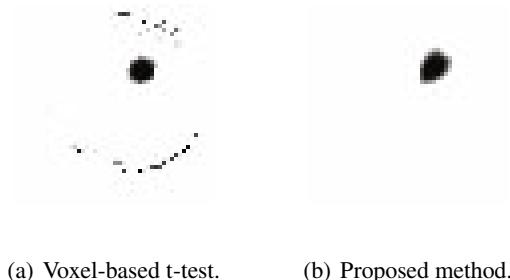


Fig. 1. Numerical Phantoms: two typical individuals from the group 1 (the two at the left) and two from group 2



(a) Voxel-based t-test. (b) Proposed method.

Fig. 2. Results using numerical phantoms

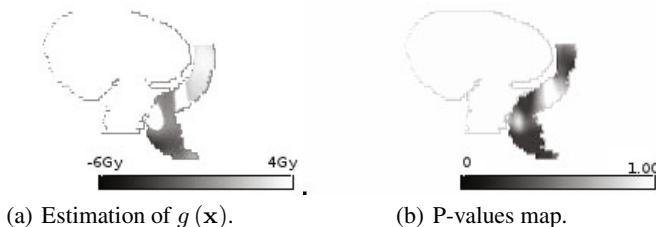


Fig. 3. Results of the rectal bleeding study with the proposed methodology

2.1 Numerical Phantoms

Two groups of numerical phantoms were created. Firstly, twenty 30mm FWHM (Full Width at Half Maximum) Gaussian images were generated representing 20 individuals. Their centres and maxima were randomly shifted to simulate misregistration effects. For 10 of them, a fixed 10 mm FWHM Gaussian pattern was added, yielding a second group for comparison. The goal of this experiment was to detect the hidden pattern characterizing the second group. Fig. 1 depicts two typical individuals from each group. A classical voxel wise two sample t-test was performed for comparison. The results of both methods are shown in Fig. 2. The method was able to accurately detect the hidden pattern in the second group. Compared with the voxel-wise two-sample t-test,

results suggest that the proposed method has a lower false positive rate (which appears around the sphere caused by the misregistration effects). Compared to the t-test, with a significance level of $\alpha = 0.05$, the proposed method outperformed the detection of true positives by 14% (voxels within the hidden pattern) and was more robust to the false positive rate (voxels outside the hidden pattern) by 40%.

2.2 Clinical Data: Predictive Model of Rectal Bleeding from Planned Dose Distribution

Within the framework of prostate cancer treatment by external beam radiotherapy, the goal of this voxel-wise analysis was to investigate the relationship between the treatment plan and the secondary effects produced by a harmful over-irradiation of the neighbouring organs at risk specifically the rectum. We selected 33 prostate cancer patients treated with external radiotherapy. Clinical outcomes (rectal bleeding, within a two years follow-up) and 3D dose distributions were available. For each patient the dose was computed on the 512x512x256 pelvic CT scans, using the manual delineations, according to the standard clinical protocol. We first mapped the organs and the dose to a common template using a hybrid organ/intensity non rigid registration method, allowing the alignment of barycentres and neighbouring structures across the population. Thus, two groups were constituted, namely patients who presented rectal bleeding within a two-year period after the treatment (17 patients) vs. those who did not (16 patients). Thus, according to the eq. (1), the planned dose distribution for the individual i , along the rectum, was modelled by the term $f_i(x)$. The underlying mean of each of the groups was $m(\mathbf{x}) + g(\mathbf{x})$ and $m(\mathbf{x})$, respectively, and the individual specific deviation from the mean behaviour was $e_i(x)$, $i = 1, \dots, 33$. The method was compared to a two-sample t-test, performed at a voxel-basis leading to the computation of three dimensional maps of the mean dose differences and the p-values. Anatomical regions where the differences were statistically significant were identified and correlated with the corresponding toxicity event. Results are shown in Fig. 3. The 3D dose difference and p-values maps suggest that there is a correlation between a higher dose delivered to the rectum and the bleeding. More importantly, the method allowed the highlighting of specific areas where the dose produced organ damage and, by controlling other clinical variables, recommendations for treatment planning could be given.

3 Conclusion

In this paper, we have presented a new method for performing voxel-based comparisons within a population, which overcomes some of the issues with classical voxel-based approaches. Because it exploits intra-individual spatial correlation at each voxel location, the method has a lower false positive rate and it can better deal with large variances between groups. Results demonstrate an improved sensitivity and reliability for group analysis compared with standard voxel-wise methods. Examples carried out on numerical phantoms and real data illustrate the benefits of this approach. We applied the methodology on clinical data, to show its feasibility, in order to explain the risks associated with spatial distribution of the dose in the case of rectal bleeding. Future

work will include the comparison with different voxel-based techniques and to extend the field of clinical applications where the voxel-based morphometry approach has been used. An important thing to note is that the method can be extended to include clinical data in order to identify more complicated patterns.

Acknowledgements. This work was partially supported by the European University of Brittany (UEB). The authors would like to thank Marian Lee for her valuable comments and help.

References

1. Friston, K.J., Holmes, A.P., et al.: Fr: Statistical Parametric Maps in Functional Imaging: A General Linear Approach. *Human Brain Mapping*, 189–210 (1995)
2. Ashburner, J., Friston, K.: Voxel-based morphometry—the methods. *Neuroimaging*. 11(6 Pt 1), 805–821 (2000)
3. Yuan, Q., Zou, L., Chen, Q.: Voxel-based morphometric study of brain structure in alzheimer's disease. *Sichuan Da Xue Xue Bao Yi Xue Ban* 39(3), 496–499 (2008)
4. Hua, X., Leow, A.D., et al.: Tensor-based morphometry as a neuroimaging biomarker for alzheimer's disease: an mri study of 676 ad, mci, and normal subjects. *Neuroimage* 43(3), 458–469 (2008)
5. Leow, A., Yanovsky, I., et al.: Statistical properties of jacobian maps and the realization of unbiased large-deformation nonlinear image registration. *IEEE Transactions on Medical Imaging* 26(6), 822–832 (2007)
6. Chételat, G., Desgranges, et al.: Direct voxel-based comparison between grey matter hypometabolism and atrophy in alzheimer's disease. *Brain* 131(Pt 1), 60–71 (2008)
7. Desgranges, B., Matuszewski, et al.: Anatomical and functional alterations in semantic dementia: a voxel-based mri and pet study. *Neurobiol. Aging* 28(12), 1904–1913 (2007)
8. Friston, K.J., Holmes, A.P., et al.: Analysis of fMRI Time-Series Revisited. *NeuroImage* 2, 45–53 (1995)
9. Kupchak, C., Battista, J., Dyk, J.V.: Experience-driven dose-volume histogram maps of NTCP risk as an aid for radiation treatment plan selection and optimization. *Med. Phys.* 35(1), 333–343 (2008)
10. Heemsbergen, W.D., Al-Mamgani, et al.: Urinary obstruction in prostate cancer patients from the dutch trial (68 gy vs. 78 gy): Relationships with local dose, acute effects, and baseline characteristics. *Int. J. Radiat. Oncol. Biol. Phys.* (January 2010)
11. Witte, M.G., Heemsbergen, W.D., et al.: Relating dose outside the prostate with freedom from failure in the dutch trial 68 gy vs. 78 gy. *Int. J. Radiat. Oncol. Biol. Phys.* 77(1), 131–138 (2010)
12. Bookstein, F.: "Voxel-based morphometry" should not be used with imperfectly registered images. *Neuroimage* 14(6), 1454–1462 (2001)
13. Ashburner, J., Friston, K.: Why voxel-based morphometry should be used. *NeuroImage* 14, 1238–1243 (2001); PMID: 11707080
14. Genovese, C.R., Lazar, N.A., Nichols, T.: Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *Neuroimage* 15(4), 870–878 (2002)
15. Nichols, T.E., Holmes, A.P.: Nonparametric permutation tests for functional neuroimaging: a primer with examples. *Hum. Brain Mapp.* 15(1), 1–25 (2002)
16. Worsley, K., Evans, A., et al.: A three dimensional statistical analysis for cbf activation studies in human brain. *J. Cereb. Blood Flow Metab.* 12, 900–918 (1992)
17. Hoover, D.R., Rice, J., et al.: Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika* 85, 809–822 (1998)

18. Wu, H., Zhan, J.T.: Nonparametric Regression Methods for Longitudinal Data Analysis. John Wiley & Sons, Inc., New Jersey (2006)
19. Chen, Y., Guo, W.: A local nonparametric model for simultaneous image segmentation and adaptive smooth. computational and applied mathematics technical report 07-34, UCLA (2007)
20. Roche, A., Mériaux, S., Keller, M., Thirion, B.: Mixed-effects statistics for group analysis in fMRI: A nonparametric maximum likelihood approach. Neuroimage 38, 501–510 (2007)
21. Tibshirani, R., Hastie, T.: Local likelihood estimation. Journal of American Statistical Association 82, 559–567 (1987)
22. R Development Core Team: R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria (2010) ISBN 3-900051-07-0
23. Pinheiro, J., Bates, D., DebRoy, S., Sarkar, D.: R Development Core Team: nlme: Linear and Nonlinear Mixed Effects Models (2010); R Package Version 3.1-97

A Machine Learning Approach to Tongue Motion Analysis in 2D Ultrasound Image Sequences

Lisa Tang¹, Ghassan Hamarneh¹, and Tim Bressmann²

¹ Medical Image Analysis Lab., School of Computing Science, Simon Fraser University
² Department of Speech-Language Pathology, Faculty of Medicine, University of Toronto

Abstract. Analysis of tongue motions as captured in dynamic ultrasound (US) images has been an important tool in speech research. Previous studies generally required semi-automatic tongue segmentations to perform data analysis. In this paper, we adopt a machine learning approach that does not require tongue segmentation. Specifically, we employ advanced normalization procedures to temporally register the US sequences using their corresponding audio files. To explicitly encode motion, we then register the image frames spatio-temporally to compute a set of deformation fields from which we construct the velocity-based and spatio-temporal gestural descriptors, where the latter explicitly encode tongue dynamics during speech. Next, making use of the recently proposed Histogram Intersection Kernel, we perform support vector machine classification to evaluate the extracted descriptors with a set of clinical measures. We applied our method to speech abnormality and tongue gestures prediction. Overall, differentiating tongue motion, as produced by patients with or without speech impediments on a dataset of 24 US sequences, was achieved with classification accuracy of 94%. When applied to another dataset of 90 US sequences for two other classification tasks, accuracies were 86% and 84%.

1 Introduction

In recent decades, ultrasound imaging research has made great strides capturing the intricate and highly coordinated nature of articulatory movements for speech. However, the tongue makes complex motion and complicated postural adjustments during speech production, making perceptual assessments of speech problems from US images extremely challenging.

Accordingly, the development of reliable and robust quantitative indicators for what constitutes normal tongue movement or what characterizes articulation errors and/or speech disorders is highly desirable. These indicators would especially be useful to inform treatment strategies for speech-language pathologists (SLP) [1] as these indicators would provide more explicit information of the tongue motion (e.g. which parts of the tongue have moved incorrectly and how incorrectly).

As a first step in developing these indicators, this paper employs a learning approach to analyze mid-sagittal tongue motion as captured in US image sequences and, in particular, seeks to develop a descriptor that can capture *spatio-temporal tongue gestures*, i.e. the dynamics of the tongue contour over time. For this task, our method employs

non-rigid registration where motion fields describing pixel displacements between consecutive frames are used to construct descriptors that encode tongue trajectories as produced during speech. We then train a support vector machine (SVM) classifier using paired data to predict a set of known measures using our proposed descriptors.

There exist several related works that examined image-captured motions for clinical problems, albeit a majority was developed specifically for cardiac motion analyses. In heart motion abnormality detection, for example, researchers have employed shape-based, tensor-based, and motion-based descriptors to study the dynamics of myocardium deformation [2]. However, direct application of these methods to tongue motion analysis is challenging due to the lack of image features for locating tongue surfaces. Due to US wave reflectance, for instance, parts of the tongue may appear missing in a frame. As 2D US sequences capture only the mid-sagittal portion of a tongue, it is often the case that the entire tongue appears as missing. Tongue motion in daily speech also does not exhibit regularity, but rather consist of intricate tongue “gymnastics”. All of these issues render reliable tongue contour detection and tracking extremely difficult.

For analysis of tongue motions from US images, most of the previous studies [3, 4, 5, 6] only examined tongue motions using velocity measurements that relied on manual delineation of tongue contours, which is both labourious and difficult [7]. There exist a few studies that employed machine learning algorithms, e.g. [8, 9], which nevertheless did not examine spatio-temporal gestures like we do and had very different applications than ours.

In contrast to the above references, our approach not only analyzes horizontal and vertical displacements, but also examines motion patterns. To develop robust descriptors, we make use of the accompanying audio files to perform temporal alignment of the US sequences. We then employ a well-validated registration algorithm [10] to reliably obtain a set of spatial correspondences that explicitly represent motion and subsequently extract a set of descriptors to capture dynamics of the tongue from the set of correspondences. In evaluating the effectiveness of the descriptors, we performed three clinically driven classification tasks. As we will show in our experimental results, our approach is capable of analyzing image sequences that contain a much wider variety of utterances than those encountered in previous studies [4, 5, 6, 11, 12].

2 Methods

Two sets of data were collected in this study, each of size N . In each study n , both audio recordings A_n and US imaging U_n were acquired while a subject recited a passage or articulated a sequence of utterances. Details on how the data is used are given in Sec. 3.

Our approach begins with the normalization of the data, which is an important step to ensure that the descriptors we extract are comparable across subjects. In our problem, the normalization step involves the selection of a set of keyframes in which the same set of utterances was articulated by different subjects. For this task, we perform temporal alignment on all N image sequences U by performing Dynamic Time Warp (DTW) on all audio recordings. With the keyframes extracted for each sequence, we then perform spatio-temporal registration to obtain a set of deformation fields from which we extract our descriptors. From the obtained deformation fields that describe tongue motions with

respect to a reference frame, we next extract a set of measurements from each field and encode these measurements with a histogram-based representation to ensure robustness against outlier displacement vectors. Lastly, we train a specialized SVM classifier to learn possible correlations between the features and a set of known clinical measures.

2.1 Audio-Based Temporal Alignment via Dynamic Time Warp

As explained earlier, normalization of the data is crucial to ensure that extracted features are meaningful and comparable across subjects. In our US data, because reading speeds varied across subjects, the frame number that corresponds to the same sound can be different across subjects. This called for temporal normalization. Only after we have resolved the temporal misalignments that exist between US sequences can we attribute the differences in our overall training data to differences in tongue gestures.

Correcting temporal misalignments based on the US images is extremely difficult as changes in appearance features between frames are very subtle. Conversely, distinctive temporal landmarks can be easily identified in the audio signals that accompany the US data. We thus used the audio signals to perform temporal registration of the US frames. Firstly, one of the sound files in each dataset is chosen as the audio template. Next, we employ the Dynamic Time Warping (DTW) algorithm of [13] (based on dynamic-programming) to temporally align each extracted sound originating from the template to those matched in the remaining sequences. To align two audio signals, DTW operates in the space of the time frames constructed with respect to the two sequences and seeks a path through this space that maximizes the local match between the aligned time frames, subject to some ordering constraints. The sum of these local similarity costs is then used to estimate how well two signals match. Temporal registration of each sequence with respect to the template consequently yields a set of time points that indicate when the same sounds were made across all subjects. Based on the frame rate of the captured US sequences, we calculate, for each sequence n , a set of k frame indices that essentially provide the temporal correspondences in U . However, converting from time to frame number I_n may result in a frame index that falls in the space between two US frames. In this case, linear interpolation on such pair of frames was done. Finally, extracting a subset from U_n based on I_n yields U_n^* that are temporally aligned across subjects.

2.2 Explicit Characterization of Dynamic Gestures via 2D+Time Registration

In [9], image features were used as descriptors to differentiate static tongue gestures. However, as our goal is to differentiate between different *dynamic* tongue gestures (spatio-temporal vs. just spatial gestures [9]), we need to capture the motion/gestural changes that occur across time. Accordingly, we performed spatio-temporal registration on each of U_n^* to explicitly find a set of inter-frame per-pixel correspondences, from which we extract information to capture tongue motions and gestural changes.

For this task, we employ the latest state-of-art registration method of Metz et al. [10] that enforces both spatial *and* temporal smoothness on sought deformation fields; the latter in particular would help decrease the sensitivity of registration to sporadic disappearances of image features of the tongue due to US wave reflectances, etc. (Fig. 1a).

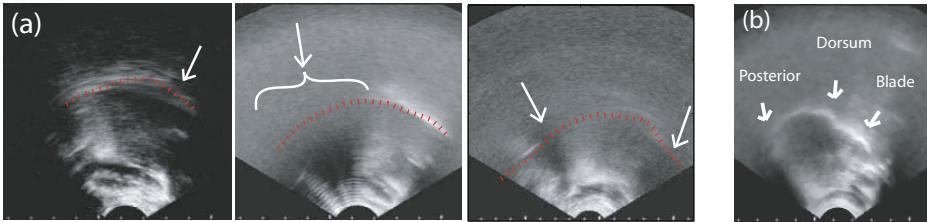


Fig. 1. (a) US frames showing missing parts of the tongue. (b) Different tongue regions.

This algorithm was designed to solve the parameters of a Lagrangian 2D+time transformation model that would minimize the intensity variance at corresponding spatial locations across time.

2.3 Feature Extraction

Spatio-temporal registration on all N US subsequences of length k results in a set of displacement vector fields $\{\mathbf{D}_{(s,1)}, \mathbf{D}_{(s,2)}, \dots, \mathbf{D}_{(s,i)}, \dots, \mathbf{D}_{(s,k-1)}\} \forall s = 1 : n$, each of which describes the coordinate mapping from frame i to frame $i + 1$. From these, we calculated a set of descriptors of two types which we detail below.

Regional velocity-based descriptors. Following [3], which concluded that velocity is an indicator for normal and abnormal tongue motion, we have built our descriptors based on velocities.

In [11], a hard threshold was applied on the motion vectors of each field such that only displacement vectors with magnitude above the mean were considered. We argue that this approach can easily discard potentially relevant information and that analysis on the remaining vectors would subsequently depend on the threshold chosen. As a better alternative, we employ histograms to encode all available information in the deformation fields. Specifically, we compute the distributions of the x - and y -components of each velocity vector to construct our histogram-based descriptors which we denote as V_x and V_y .

Furthermore, it is of clinical interest to examine whether differences in features at different parts of the tongue exist as this may give implications to impediments of specific muscles. Therefore, we have regionalized the encoding of features by extracting the above features in terms of the left, middle, and right sides of the image domain, which corresponds to regions of the tongue blade, dorsum and posterior (Fig. 2b). For brevity, we append each feature type with a suffix ‘B’, ‘D’, or ‘P’ (e.g. V_x -D) to denote the blade, dorsum, and posterior, respectively. Each feature type was encoded with $h = 20$ histogram bins, where h was chosen based on initial experiments.

Spatio-temporal gestural descriptors. We have developed these descriptors to explicitly encode changes in motion over time. Rather than treating each displacement vector in a field as a feature component, which would involve a feature vector of size equal to the product of the image dimensions, d , we perform dimensionality reduction to encode the most relevant information in each field. To compactly represent the obtained displacement vector field of each frame, we employ principal component analysis (PCA)

to project the displacement fields onto a feature space from which we obtain a finite set of orthogonal principal components that would constitute, up to a certain accuracy, a subspace for the representations of the most likely configurations of x and y displacements. Subsequently, the x and y components in each field are represented compactly in terms of their projections onto the computed set of principal components. Note that while this dimensionality reduction step is similar to the EigenTongue feature space of [8], our approach operates on the displacement vector fields, rather than the original static US image frames of the tongue. For brevity, we denote the principal coefficients describing the projections of frame i as $\mathbf{P}_x^i, \mathbf{P}_y^i$ for x and y dimension, respectively. With the principal coefficients representing each field computed, we then concatenate these coefficients of all frames in each US subsequence to form our gestural descriptor, i.e. $[\mathbf{P}_x^1 \mathbf{P}_y^1 \dots \mathbf{P}_x^k \mathbf{P}_y^k]$. Note that we use the first $p << d$ principal component coefficients corresponding to the first p principal components that account for 98% of the variance in all vector fields.

2.4 Training the SVM Classifier

Our next step is to train a SVM classifier to build a model that correlates the extracted features to a set of given measures (to be given in Sec. 3). Prior to training and classification, each component of all feature vectors is normalized to $[0,1]$.

In SVM, the classifier simultaneously minimizes the classification error and maximizes the margin between classes [14]. In measuring distances between our descriptors, we employ the Histogram Intersection Kernel (HIK) that was recently proposed by Wu [14]. Aside from being computationally efficient, it was empirically shown to be insensitive to the C parameter in SVM and is capable of achieving a higher accuracy in SVM classification than linear or radial basis function kernels in various application domains [14]. Under this kernel, the distance between two vectors a and $b \in \mathbb{R}^r$ is defined as $\mathcal{K}(a, b) = \sum_{j=1}^r \min(a_j, b_j)$.

In training the SVM classifier, we employ the deterministic Intersection Coordinate Descent (ICD) algorithm that was also proposed by Wu [14], which solves the classification problem without re-encoding the input data and explicitly finds the feature space decision boundary.

3 Experiments

Accuracy of tongue tracking directly affects the sensitivity of the extracted features for our learning task. It is therefore important to assess the accuracy of the spatio-temporal registration results. To access accuracy, 12 randomly selected registration results were validated. From each sequence, two sets of manually created tongue segmentations were acquired and subsequently treated as bronze-standard. Segmentation of the first keyframe was then selected as the template and was ‘propagated’ to the subsequent frames using the obtained deformation fields. We then computed the mean distances between the propagated template with the bronze-standard segmentations. Fig. 2a shows the distribution of the mean Euclidean distances between the propagated and bronze-standard contours in all segmented frames. Overall, the mean Euclidean

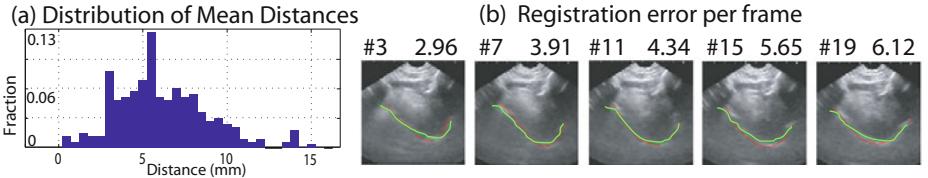


Fig. 2. Evaluation of registration results. Distribution of (a) mean distances between the obtained and expert-delineated contours in all segmented frames. (b) One registration result showing expert-delineated tongue contours (in green) and those obtained via registration (in red). Error (in mm) noted on top of each image, after each frame number.

distance was 5.82 mm. This is acceptable as it compares well with the inter-operator segmentation variability of 5.3 mm, which we measured based on segmentation results from two operators. Fig. 2b shows an example of the propagated and bronze-standard segmentations.

Having validated the registration results, we then extracted the proposed features from the obtained fields. Next, we performed two experiments to evaluate our approach for tongue motion analysis. The first experiment examines how well our velocity-based descriptors can predict normal vs. abnormal speech. The second experiment questions whether different tongue motions can be effectively encoded with our gestural descriptors to such an extent that they can be used to predict different tongue gestures.

Differentiating normal from abnormal. The first experiment examined how our velocity-based descriptors can differentiate tongue motions produced from patients with and without speech impediments. For this experiment, we employed the data from [3], which consists of 12 studies from patients with normal speech and 12 studies from the same patients with abnormal speech (before and after undergoing lateral partial glossectomies). Each study consists of an audio recording and an US image sequence of 200–420 frames that captures the patient’s tongue motion while reading a passage.

For motion analysis, we extracted $k = 120$ keyframes from each of the 24 US image sequences, thus giving rise to a set of $k \times 24$ frame samples of tongue motion, each associated with a label of ‘normal’ or ‘abnormal’. We then performed temporal registration on the extracted keyframes for each patient study to obtain a set of displacement vector fields from which we extracted the features described in Sec. 2.3. Next, we trained a SVM-classifier to build a model that would predict the type of speech impediments from the set of extracted features.

We now report the ‘with vs. without speech impediments’ mean classification accuracies¹ of individual features as obtained from a 10-fold cross validation experiment. For the velocity-based features, these were 86, 90, 81, 89, 80, and 81% for V_x -B, V_y -B, V_x -D, V_y -D, V_x -P, and V_y -P respectively. This result is in agreement with the results obtained in [3], which showed that velocity was an indicator for abnormal speech, and

¹ For each classification task, we ran 12 repeated trials, each with a different value of C (parameter in SVM) within $[2^{-8}, 2^8]$ and computed accuracy for each trial as the fraction of correctly classified samples over all samples. We then report accuracy as the mean over all 12 trials.

is an encouraging one as it reflects that our approach gave results identical to a method that relied on manual segmentation. Furthermore, the mean classification accuracy of Vy-B was 9% greater than the accuracy of Vy-P, suggesting that the *y*-velocity exhibited in the blade region differentiated between normal and abnormal tongue motions much more effectively than those exhibited in the posterior region. Using all velocity-based features simultaneously, we achieved mean classification accuracy of 94%.

Differentiating utterance types using the gestural descriptors. Our next task examined whether the gestural descriptors can describe different motion patterns as produced by articulations of different utterances. We have done so by quantifying how well they can be used to predict 3 tongue gestures produced by articulations of 3 distinct utterances.

The data for this experiment originated from [15] and involved 9 subjects articulating three vowel-consonant-vowel (VCV) sequences. These are /aka/, /ishi/, /ushu/. Generally, articulation of each VCV sequence by all subjects involved different spatio-temporal tongue gestures, with /aka/ having the most pronounced lingual excursion, while /ishi/ and /ushu/ have relatively much less movements. In fact, from inspection of the US images, the tongue gestures of the latter two appeared similar, making the gesture-prediction task challenging.

Each subject recited the same VCV sequence 10 times consecutively before proceeding to the next sequence, yielding a set of 9×10 motion samples. After performing the normalization step outlined in Sec. 2, each motion sample consisted of $k = 30$ frames. We then extracted the gestural descriptors and applied them to 4 gesture-prediction tasks: a 3-class classification task and three binary classification tasks (/aka/ vs. /ishi/, /ishi/ vs. /ushu/, and /aka/ vs. /ushu/). Mean classification accuracies for the 4 respective tasks were 74, 86, 86, and 84%. While classification accuracy for the 3-class problem was suboptimal, all binary classifications gave comparable performances, suggesting that the effectiveness of the gestural features persisted even with the subtle differences between similar tongue gestures.

Differentiating normal from abnormal using the gestural descriptors. Our last experiment examined how well our SVM classifier can be trained to predict abnormal and normal tongue motion using spatio-temporal features. For this task, we collected additional data for the same set of VCV sequences that indicated whether the sequence was made with normal or abnormal speech. Using the same procedure as done in the second experiment, we performed 3 binary classification tasks, one for each of the above 3 VCV sequences. The obtained mean classification accuracies were 84, 84 and 86%, for /aka/, /ishi/ and /ushu/, respectively.

In summary, both velocity-based and spatio-temporal descriptors were shown to be capable of discriminating tongue motions as produced from subjects with and without speech impediments. We also found that different localized velocity-based features gave different levels of classification accuracies, suggesting that regional differences in tongue motion exist.

4 Conclusions

We have presented a machine learning approach to analyze and describe motions of the human tongue as captured from dynamic US image sequences. Experimental results show that our approach of extracting and encoding the proposed descriptors from spatio-temporally aligned data is effective for automatic analysis of tongue motion as captured in US sequences. Therefore, if different levels of speech impediments lead to varying levels of changes in tongue gestures, our method may be able to differentiate between gestures of *different levels* of speech impediments. This would ultimately lead to the development of a set of indicators for different types of articulation disorders. We are currently pursuing the collection of such data to test this hypothesis.

References

1. Bressmann, T.: Ultrasound imaging and its application in speech-language pathology and speech science. *Amer. Speech-Language-Hearing Assoc. Newslett.* 33(4), 204–211 (2007)
2. Punithakumar, K., Ayed, I.B., Ross, I.G., Islam, A., Chong, J., Li, S.: Detection of left ventricular motion abnormality via information measures and bayesian filtering. *IEEE Trans. Inf. Technol. Biomed.* 14, 1106–1113 (2010)
3. Rastadmeir, O., Bressmann, T., Smyth, R., Irish, J.: Increased midsagittal tongue velocity as indication of articulatory compensation in patients with lateral partial glossectomies. *J. Head and Neck* 30, 718–726 (2008)
4. Kocjancic, T.: Ultrasound study of tongue movements in childhood apraxia of speech. In: Ultrafest V, pp. 1–2 (2010)
5. Stern, M., Hagan, J., Park, J., Traub, D.: The effects of age on tongue motion and speech duration. In: Ultrafest V, pp. 1–2 (2010)
6. Shadle, C.H., Iskarous, K., Proctor, M.I.: Use of ultrasound to study differences in the tongue dorsum of voiced vs. voiceless fricatives. In: Ultrafest V, pp. 1–2 (2010)
7. Tang, L., Hamarneh, G.: Graph-based tracking of the tongue contour in ultrasound sequences with adaptive temporal regularization. In: MMBIA, pp. 1–8 (2010)
8. Hueber, T., Aversano, G., Chollet, G., Denby, B., Dreyfus, G., Oussar, Y., Roussel, P., Stone, M.: Eigentongue feature extraction for an ultrasound-based silent speech interface. In: IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2007, vol. 1, pp. I-1245–I-1248 (2007)
9. Berry, J., Diana Archangeli, I.F.: Automatic classification of tongue gestures in ultrasound images. In: Laboratory Phonology, vol. 12 (2010)
10. Metz, C., Klein, S., Schaap, M., van Walsum, T., Niessen, W.: Nonrigid registration of dynamic medical imaging data using nD+t B-splines and a groupwise optimization approach. *Medical Image Analysis* 15, 238–249 (2011)
11. Bird, S., Leonard, J., Moisik, S.: A motion vector analysis of tongue motion in SENCOFEN /qV/ and /Vq/ sequences. In: Ultrafest V, pp. 1–2 (2010)
12. Moisik, S.R.: Laryngeal Ultrasound Assessment of Retracted and Constricted Articulations by Phoneticians. In: Ultrafest V, pp. 1–2 (2010)
13. Turetsky, R., Ellis, D.: Ground-truth transcriptions of real music from force-aligned midi syntheses. In: 4th ISMIR, pp. 135–141 (2003)
14. Wu, J.: A fast dual method for HIK SVM learning. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) *ECCV 2010. LNCS*, vol. 6312, pp. 552–565. Springer, Heidelberg (2010)
15. Herold, B., Bressmann, T., Quintero, J., Hielscher-Fastabend, M., Stenneken, P., Irish, J.: Analysis of vowel-consonant-vowel sequences in patients with partial glossectomies using 2D ultrasound imaging. In: Ultrafest V, pp. 1–2 (2010)

Random Forest-Based Manifold Learning for Classification of Imaging Data in Dementia

Katherine R. Gray^{1,*}, Paul Aljabar¹, Rolf A. Heckemann^{2,3},
Alexander Hammers^{2,3}, and Daniel Rueckert¹

¹ Department of Computing, Imperial College London, United Kingdom

² Fondation Neurodis, CERMEP-Imagerie du Vivant, Lyon, France

³ Faculty of Medicine, Imperial College London, United Kingdom

katherine.gray03@imperial.ac.uk

Abstract. Neurodegenerative disorders are characterized by changes in multiple biomarkers, which may provide complementary information for diagnosis and prognosis. We present a framework in which proximities derived from random forests are used to learn a low-dimensional manifold from labelled training data and then to infer the clinical labels of test data mapped to this space. The proposed method facilitates the combination of embeddings from multiple datasets, resulting in the generation of a joint embedding that simultaneously encodes information about all the available features. It is possible to combine different types of data without additional processing, and we demonstrate this key feature by application to voxel-based FDG-PET and region-based MR imaging data from the ADNI study. Classification based on the joint embedding coordinates out-performs classification based on either modality alone. Results are impressive compared with other state-of-the-art machine learning techniques applied to multi-modality imaging data.

1 Introduction

There is much interest in the application of manifold learning techniques to medical imaging data. These techniques aim to convert high-dimensional data to a lower-dimensional representation in which further analysis, such as classification, may be more easily performed. This is relevant to neuroimaging, where a considerable amount of research focuses on the identification of imaging biomarkers which are desirable for improved diagnosis and monitoring, and drug discovery.

Typically, measures encoding pairwise similarities between images are used in manifold learning. For example, Laplacian eigenmaps [1] have been used to generate an embedding of brain MR images based on similarities derived from overlaps of their structural segmentations [2]. Laplacian eigenmaps have also

* This project is partially funded under the 7th Framework Programme by the European Commission (<http://cordis.europa.eu/ist>). Imaging data were provided by the Alzheimer's Disease Neuroimaging Initiative. KRG received a studentship from the EPSRC. RAH was supported by a research grant from the Dunhill Medical Trust.

been applied to ultrasound data to aid detection of breathing motion for image-based respiratory gating [3]. The manifold structure of brain MR images has been estimated by applying Isomap [4] using distance measures based on non-rigid transformations between image pairs [5]. A framework for fusing manifold learning steps, based on multiple pairwise similarity measures, is proposed in [6].

Here, we derive a low-dimensional manifold from labelled training images and use it to infer the clinical labels of test images mapped to this space. We use proximity measures derived from random forests [7], coupled with multidimensional scaling (MDS) [8], to learn the manifold on which to perform classification. The use of MDS with a random forest-derived proximity matrix is commonly used for low-dimensional data visualization [9]. Random forest proximities have also been successfully applied in unsupervised clustering tasks, in particular those involving genetic data [10]. We derive a supervised similarity measure which should generate a manifold that is optimal for the task of clinical group discrimination. We evaluate the proposed method using FDG-PET and MR imaging data from 287 participants in the Alzheimer’s Disease Neuroimaging Initiative (ADNI)¹.

The proposed framework facilitates the incorporation of multi-modality data, since combining proximities derived from two modalities generates an embedding that simultaneously encodes information about both. This is a key feature of the method, since neurodegenerative disorders such as Alzheimer’s disease are characterized by changes in multiple biomarkers which may prove more powerful when used in combination. Recently, improvements in classification accuracy have been reported when combining FDG-PET and MR imaging data using multi-kernel learning [11,12]. Importantly, the proposed method can combine different types of features without additional processing. We demonstrate this by application to voxel-based FDG-PET and region-based MR imaging data.

Our contributions are (1) the application of manifold learning to a diverse multi-modality set of brain images, (2) the use of a supervised proximity matrix derived from random forests to develop a clinically relevant manifold for classification, and (3) the proposal of a framework within which to combine multi-modality data in a single joint manifold, leading to state-of-the-art classification performance superior to that achievable using data from a single modality.

2 Methods

Features were extracted from the images using voxel-based analysis for the FDG-PET and region-based analysis for the MRI. Details of the image processing and feature extraction steps are provided in Section 3.1. A random forest classifier was applied to the data from each modality independently, firstly to obtain a baseline classification rate for comparison, but also to derive the proximity matrices required for manifold learning. Classical MDS was applied to each proximity matrix to generate a low-dimensional embedding for each modality. The two proximity matrices were then combined, and MDS applied to generate a joint embedding. A schematic illustration of the approach is provided in Figure 1.

¹ www.loni.ucla.edu/ADNI

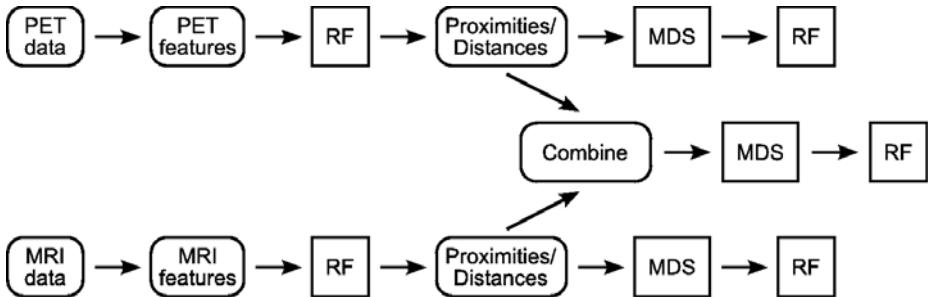


Fig. 1. An overview of the proposed analysis pipeline. Each random forest (RF) step provides a classification result whose performance is reported in Section 3.2.

2.1 Random Forests for Classification

A random forest is an ensemble classifier consisting of many decision trees, where the final predicted class for a test object is the mode of the predictions of all individual trees. For a dataset consisting of N objects, each with M features, a value $m << M$ is selected, and each tree grown as follows. To construct the training set for a tree, N objects are sampled at random with replacement. At each node in the tree, m features are randomly selected from the available M , and the node is partitioned using the best possible binary split. Each tree is fully grown without pruning. We use the R implementation of random forests².

2.2 Manifold Learning Based on Random Forest Proximities

Random forest proximities provide pairwise measures of the similarities between objects in the dataset. All N objects are passed down each tree in the forest, and if objects i and j finish in the same terminal node, their proximity, p_{ij} , is increased by one. The proximities are normalized by the total number of trees.

The proximities form an $N \times N$ matrix, P_{ij} , for which values of the corresponding distance matrix, $D_{ij} = 1 - P_{ij}$, may be viewed as squared distances in a Euclidean space of dimension not greater than the number of objects [8]. MDS can be applied to this distance matrix to generate a lower-dimensional embedding for the data. To generate a joint embedding that simultaneously incorporates information from both modalities, the distance matrices derived from the individual modalities are additively combined, and MDS applied to the resulting joint proximity matrix.

3 Data and Results

3.1 Imaging Data and Feature Extraction

We applied the proposed methods to images of 287 ADNI participants for whom baseline 1.5T MRI and FDG-PET images are available. These include 71 patients

² cran.r-project.org/web/packages/randomForest

with Alzheimer’s disease (AD), 147 patients with mild cognitive impairment (MCI), and 69 age-matched healthy controls (HC).

MRI: region-based feature extraction. Automatic whole-brain segmentations into 83 anatomical regions were prepared in native-space using multi-atlas propagation with enhanced registration (MAPER), an approach that has been previously described and validated for use in AD [13]. The segmentations are available to download through the ADNI website, and full details of the procedure and morphometric analysis are presented in [14]. Individual tissue probability maps for gray matter, white matter and cerebrospinal fluid were obtained using FSL FAST³. For feature extraction, masked segmentations were employed, in which all regions except ventricles, central structures, cerebellum and brain-stem were masked with a gray matter label, and lateral ventricles with a cerebrospinal fluid label. Regional volumes were calculated and normalized by the intracranial volume, resulting in 83 volumetric region-based features per image.

FDG-PET: voxel-based feature extraction. Each FDG-PET image was motion-corrected as necessary, converted to a 30-minute static image, and affinely aligned with the corresponding MRI using the Image Registration Toolkit (IRTK)⁴. An affine transformation was preferred over a rigid one because it can account for any scaling or voxel size errors remaining after phantom correction of the MRI [15]. Using the “Segment” module of SPM5⁵, each MRI was linearly and non-linearly deformed to the MNI template. The resulting transformation parameters were applied to the MR-space FDG-PET images using a trilinear interpolation. The MNI-space FDG-PET images were smoothed to a common isotropic spatial resolution of 8mm FWHM using scanner-specific kernels [16], and then by an additional 8mm FWHM isotropic Gaussian kernel. The smoothed images were normalized to account for inter-subject variability in overall radioactivity using a reference cluster derived from an independent dataset [17]. A MNI-space brainmask was applied to each normalized FDG-PET image, thresholded at 50% to exclude background, and signal intensities were extracted from each voxel, resulting in 239,304 features per image.

3.2 Results

We assess classification performance between two clinically significant pairs of groups: AD patients vs HC, and MCI patients vs HC. All experiments were assessed using ten-fold cross-validation. Since there are approximately twice as many MCI patients as HC, these data were balanced according to the class distributions during training. Before classification, the number of features randomly selected at each node, m , and the number of trees grown in each forest, t , had to be selected. We used 5,000 trees, and the recommended default value, $m = \sqrt{M}$.

³ www.fmrib.ox.ac.uk/fsl

⁴ www.doc.ic.ac.uk/~dr/software

⁵ www.fil.ion.ucl.ac.uk/spm

Table 1. Classification accuracy, sensitivity and specificity following application of a random forest classifier to the original imaging data. Classification is performed using ten-fold cross-validation, and mean (standard error) values are reported.

	AD vs HC			MCI vs HC		
	Acc. (%)	Sens. (%)	Spec. (%)	Acc. (%)	Sens. (%)	Spec. (%)
Region-based MRI	84.4 (1.6)	83.2 (3.5)	85.5 (3.0)	64.4 (3.0)	59.3 (4.5)	75.7 (5.2)
Voxel-based FDG-PET	87.9 (2.6)	92.0 (3.9)	83.8 (4.7)	63.9 (2.3)	59.1 (2.8)	74.1 (4.6)
Concatenated features	87.9 (2.6)	92.0 (3.9)	83.8 (4.7)	64.3 (2.4)	59.8 (3.2)	74.0 (4.6)

Table 2. Classification accuracy, sensitivity and specificity obtained following the application of a random forest classifier to the embedded data from the two individual modalities, as well as the joint embedding. Classification is performed using ten-fold cross-validation, and mean (standard error) values are reported.

	AD vs HC			MCI vs HC		
	Acc. (%)	Sens. (%)	Spec. (%)	Acc. (%)	Sens. (%)	Spec. (%)
Region-based MRI	87.2 (2.0)	87.5 (3.2)	86.9 (2.6)	64.8 (3.0)	64.8 (3.9)	65.5 (6.7)
Voxel-based FDG-PET	87.8 (2.6)	91.8 (2.9)	83.8 (5.1)	65.3 (1.9)	65.3 (2.9)	65.2 (3.8)
Combined embedding	90.0 (2.6)	88.9 (3.4)	89.8 (3.8)	75.5 (2.2)	76.9 (3.2)	72.4 (4.5)

Table 1 shows the classification performance following the application of a random forest classifier to the original imaging data. In addition, the FDG-PET and MRI features were combined by concatenation, and a random forest classifier applied. The performance based on this feature set is also shown, and does not significantly differ from the results based solely on the FDG-PET features.

After applying MDS to the proximity matrix for each modality, the eigenvectors corresponding to the 25 largest-valued eigenvalues were used in generating low-dimensional embeddings. The value of 25 was empirically determined to ensure that zero-valued eigenvalues were not included, whilst capturing the maximum possible amount of information. Table 2 shows the classification performance following application of a random forest classifier to the separate embedding coordinates for each modality. The accuracy achieved based on the embedding coordinates does not differ significantly from that achieved using the original imaging data. Results based on the joint embedding are also shown. This out-performs the corresponding application to the separate embedding coordinates. The improvement is significant ($p < 0.05$) for the MCI vs HC classification.

It is possible to extract estimates of the importance of each feature for classification from the random forest. The importance measure for an individual feature is determined by summing the decreases in the Gini impurity criterion [18] over all nodes in the forest that are partitioned based on that feature. The Gini impurity of a node measures the likelihood that an object would be incorrectly labelled if it were randomly assigned a label according to the distribution of labels within the node. Feature importances for discriminating between clinical

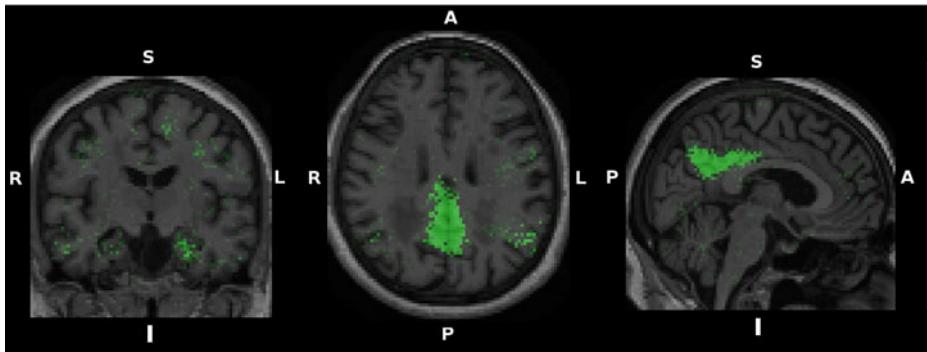


Fig. 2. Feature importances for discriminating between AD patients and HC using FDG-PET imaging data. All voxels with an importance measure greater than 5% of the maximum value are displayed in green, and overlaid onto the MNI-space MR image of a typical HC. R: right, L: left, S: superior, I: inferior, A: anterior, P: posterior.

groups were extracted from the random forest classifiers applied to the original imaging data. For FDG-PET, Figure 2 shows the locations of the most important voxels for discriminating between AD patients and HC. Important voxels are mostly found in the posterior cingulate gyrus, parietal lobe, posterior temporal lobe, and around the hippocampus. For discriminating between MCI patients and HC, important voxels are also located in these areas, but at lower magnitude, as well as in the frontal cortex. For MRI, the most important features include the hippocampus, amygdala, and other medial temporal lobe structures. A larger number of regions are identified as relatively important in distinguishing MCI patients from HC, but with lower magnitude.

4 Discussion and Conclusions

We have presented a framework in which proximity measures derived from random forests are used to learn a low-dimensional manifold from labelled training data and then to infer the clinical labels of test data mapped to this space. The applicability of the method to a large and diverse set of brain images is demonstrated using FDG-PET and MR images from the ADNI study. A key feature of the method is its ability to incorporate multi-modality data, since it is unlikely that AD can be fully characterized by a single biomarker. The method could be extended to include any number of features, for example longitudinal imaging data, non-imaging data, and categorical and genetic features. We demonstrate the potential to combine different types of features without additional processing, providing the freedom to perform feature extraction in the most appropriate way for each individual dataset. In other similar methods, such as that described in [6], a scaling step is required before separate embeddings may be combined.

We assess the performance of the method by application to voxel-based FDG-PET and region-based MR imaging data from the ADNI study. Classification

based on the joint embedding derived from these two modalities out-performs classification based on either modality alone. This supports previous suggestions that there is some complementary information between MRI and FDG-PET which can be exploited to produce a more powerful combined biomarker for AD. The lack of significant difference between classification performance based on the original imaging data compared with the embedding coordinates for each individual modality is to be expected, since a random forest is already a nonlinear classifier. The motivation for the embedding step was the incorporation of multi-modality data. We demonstrate that a simple concatenation of FDG-PET and MRI features does not optimally combine these data, as this does not improve classification performance compared with the single modalities.

The classification accuracy for discriminating MCI patients from HC based on the joint embedding is impressive compared with other state-of-the-art multi-modality machine learning techniques. For example, [11] also report an accuracy of 76% using multi-kernel learning, but based on the combination of MRI, FDG-PET and cerebrospinal fluid biomarkers. The classification accuracy for discriminating AD patients from HC is in line with other state-of-the-art methods which use either single-modality [21] or multi-modality [11,12] imaging data. Since the diagnostic labels used as a gold-standard themselves have an accuracy of around 90% [22], significant further improvement may not be possible using this dataset.

The ability to extract feature importances from the random forest is valuable because it allows verification that the features contributing most to the classification make biological sense. The most important features for discriminating between HC and both AD and MCI patients are in line with those known to be different between clinical groups in the respective modalities [19,20]. The important features for discriminating AD patients from HC are more localized to affected areas, with the more challenging discrimination between MCI patients and HC also requiring features spread across a wider area of the brain.

The method is readily generalizable, in that the manifold learning step could be performed using Laplacian eigenmaps, Isomap, or any other suitable algorithm. Classification performance could be assessed by applying a clustering algorithm to the embedding coordinates, and proximities could be combined using a more sophisticated metric. Future work will explore these options, and consider other promising datasets, such as cerebrospinal fluid biomarkers.

References

1. Belkin, M., Niyogi, P.: Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation* 15(6), 1373–1396 (2003)
2. Aljabar, P., Rueckert, D., Crum, W.: Automated morphological analysis of magnetic resonance brain imaging using spectral analysis. *Neuroimage* 43(2), 225–235 (2008)
3. Wachinger, C., Yigitsoy, M., Navab, N.: Manifold learning for image-based breathing gating with application to 4D ultrasound. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6362, pp. 26–33. Springer, Heidelberg (2010)

4. Tenenbaum, J., de Silva, V., Langford, J.: A global geometric framework for non-linear dimensionality reduction. *Science* 290(5500), 2319–2323 (2000)
5. Gerber, S., Tasdizen, T., Joshi, S., Whitaker, R.: On the manifold structure of the space of brain images. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5761, pp. 305–312. Springer, Heidelberg (2009)
6. Aljabar, P., Wolz, R., Srinivasan, L., Counsell, S., Boardman, J.P., Murgasova, M., Doria, V., Rutherford, M.A., Edwards, A.D., Hajnal, J.V., Rueckert, D.: Combining morphological information in a manifold learning framework: Application to neonatal MRI. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6363, pp. 1–8. Springer, Heidelberg (2010)
7. Breiman, L.: Random forests. *Machine Learning* 45(1), 5–32 (2001)
8. Cox, T.F., Cox, M.A.A.: Multidimensional scaling. Chapman and Hall, Boca Raton (2001)
9. Hastie, T., Tibshirani, R., Friedman, J.: The elements of statistical learning: data mining, inference, and prediction, 2nd edn. Springer, Heidelberg (2011); corrected 5th printing
10. Shi, T., Horvath, S.: Unsupervised learning with random forest predictors. *J. Comp. Graph. Stat.* 15(1), 118–138 (2006)
11. Zhang, D., Wang, Y., Zhou, L., et al.: Multimodal classification of Alzheimer's disease and mild cognitive impairment. *Neuroimage* 55(3), 856–867 (2011)
12. Hinrichs, C., Singh, V., Xu, G., et al.: Predictive markers for AD in a multimodality framework: an analysis of MCI progression in the ADNI population. *Neuroimage* 55(2), 574–589 (2011)
13. Heckemann, R.A., Keihaninejad, S., Aljabar, P., et al.: Improving intersubject image registration using tissue-class information benefits robustness and accuracy of multi-atlas based anatomical segmentation. *Neuroimage* 51(1), 221–227 (2010)
14. Heckemann, R.A., Keihaninejad, S., Aljabar, P., et al.: Automatic morphometry in Alzheimer's disease and mild cognitive impairment. *Neuroimage* 56(4), 2024–2037 (2011)
15. Clarkson, M.J., Ourselin, S., Nielsen, C., et al.: Comparison of phantom and registration scaling corrections using the ADNI cohort. *Neuroimage* 47(4), 1506–1513 (2009)
16. Joshi, A., Koeppe, R.A., Fessler, J.A.: Reducing between scanner differences in multi-center PET studies. *Neuroimage* 46(1), 154–159 (2009)
17. Yakushev, I., Hammers, A., Fellgiebel, A., et al.: SPM-based count normalization provides excellent discrimination of mild Alzheimer's disease and amnestic mild cognitive impairment from healthy aging. *Neuroimage* 44(1), 43–50 (2009)
18. Breiman, L., Friedman, J.H., Olshen, R.A., et al.: Classification and regression trees. Wadsworth, Belmont (1984)
19. Hampel, H., Burger, K., Teipel, S.J., et al.: Core candidate neurochemical and imaging biomarkers of Alzheimer's disease. *Alzh. & Dementia* 4(1), 38–48 (2008)
20. Patwardhan, M.B., McCrory, D.C., Matchar, D.B., et al.: Alzheimer disease: operating characteristics of PET – a meta-analysis. *Radiology* 231(1), 73–80 (2004)
21. Cuingnet, R., Gerardin, E., Tessieras, J., et al.: Automatic classification of patients with Alzheimer's disease from structural MRI: a comparison of ten methods using the ADNI database. *Neuroimage* 56(2), 766–781 (2011)
22. Ranginwala, N.A., Hynan, L.S., Weiner, M.F., et al.: Clinical criteria for the diagnosis of Alzheimer disease: still good after all these years. *Am. J. Geriat. Psychiatry* 16(5), 384–388 (2008)

Probabilistic Graphical Model of SPECT/MRI

Stefano Pedemonte¹, Alexandre Bousse², Brian F. Hutton², Simon Arridge¹,
and Sébastien Ourselin¹

¹ The Centre for Medial Image Computing, UCL, London, United Kingdom

² Institute of Nuclear Medicine, UCL Hospitals NHS Trust, London, United Kingdom

Abstract. The combination of PET and SPECT with MRI is an area of active research at present time and will enable new biological and pathological analysis tools for clinical applications and pre-clinical research. Image processing and reconstruction in multi-modal PET/MRI and SPECT/MRI poses new algorithmic and computational challenges. We investigate the use of Probabilistic Graphical Models (PGM) to construct a system model and to factorize the complex joint distribution that arises from the combination of the two imaging systems. A joint generative system model based on finite mixtures is proposed and the structural properties of the associated PGM are addressed in order to obtain an iterative algorithm for estimation of activity and multi-modal segmentation. In a SPECT/MRI digital phantom study, the proposed algorithm outperforms a well established method for multi-modal activity estimation in terms of bias/variance characteristics and identification of lesions.

Keywords: Molecular Imaging, Emission Tomography, Multi-modality, Bayesian Networks.

1 Introduction

Algorithms for stochastic image reconstruction in Emission Tomography (PET and SPECT) are widespread in research and clinical applications for the accuracy they can provide by taking into account photon count statistics and detailed system models. Common formulations of such algorithms rely on iterative procedures in order to find an approximation of the spatial distribution of radiopharmaceutical activity that is most likely to have generated the detected photon interaction events. Low photon count and approximated system models heavily limit the resolution in emission tomographic imaging; many publications have focused on multi-modality enhanced reconstruction, where information from an intra-patient anatomical image (CT, MRI) improves the estimate of activity. Such methods are based on the assumption that activity is related to the underlying anatomy, which is linked to the CT and MRI image intensity. Methods in the literature fall within three main categories: methods that favor a piecewise uniform reconstruction by segmenting the anatomical image and subsequently applying a smoothing prior within each identified region; methods that explicitly extract boundary information from the anatomical image and relax the effect of

a global smoothing prior across the identified edges; methods based on information theoretic similarity functionals [1].

The last category has proven particularly interesting, as the involved functionals do not require either explicit segmentation nor boundary extraction from the anatomical image, steps that are inherently sensitive to noise because of selection of high frequency components of the image. The introduction of anatomical prior information via these functionals has been shown to improve the *a posteriori* estimate of activity, by reducing its *bias* and sensitivity to noise [1]. However such methods are uninformed about the imaging processes and present intrinsic problems due to the existence of multiple solutions, which determine estimates of the activity inconsistent with boundary information in the anatomical image.

In this work we develop a joint generative model that captures the complex interactions of high dimensional variables in a combined system for Emission Tomography and MRI. The interdependence of the two imaging modalities is explained by the existence of a hidden state.

Similar generative models based on a hidden state have been adopted recently in multi-modal imaging by Venkataraman *et al.* for joint estimation of brain connectivity from functional MRI (fMRI) and diffusion weighted imaging (DWI) [4] and by Hiltunen *et al.* in single-modality diffuse optical tomography for combined reconstruction-classification [5]. Earlier work on maximum a posteriori joint estimation of function and anatomy was developed by Sastry and Carson [6] who introduced a tissue composition model based on a finite mixture and Rangarajan *et al.* [7] who introduced an iterative scheme for the same model, based on maximization of mutual information.

2 Methods

Modeling the system with a PGM allows us to obtain an iterative algorithm for estimation of activity from the factorization of the joint probability distribution associated to the graphical model [3]. In the following a PGM for the multi-modal system is obtained by combining two models of the separate modalities by the use of a latent anatomical/functional state.

2.1 Probabilistic Graphical Model of SPECT

Let the radio-pharmaceutical activity within the region of interest of the patient's body be a continuous function denoted by \tilde{y} . In order to readily discretise the reconstruction algorithm, it is convenient to imagine that the activity is in the first place discrete in space. Let us approximate \tilde{y} by a set of point sources $y = y_b, b \in \{1, \dots, N_b\}$ displaced on a regular grid.

Given that each point source emits radiation at an average rate y_b proportional to the local density of radio-tracer and emission events in a same voxel are not time correlated, the number of emissions in the unit time from within voxel b is a Poisson distribution of expected value y_b . The geometry of the system and attenuation in the patient determine the probability p_{bd} that a photon emitted

in b is detected at detector pixel d (assuming binned detection). From the sum property of the Poisson distribution, the photon count in d is Poisson distributed with expected value $\sum_b p_{bd}y_b$. Given activity y , the probability to observe counts z_d in d is:

$$p(z_d|y) = \mathcal{P}\left(\sum_b p_{bd}y_b, z_d\right) \quad (1)$$

As activity determines counts, counts in each of the detector bins d are independent conditionally to activity, as expressed by the directed acyclical graph (DAG) in figure 1-A (and by the Global Markov properties if its moralized graph on the right). Given activity y , the probability to observe z is then:

$$p(z|y) = \prod_{d=1}^{N_d} \mathcal{P}\left(\sum_b p_{bd}y_b, z_d\right) \quad (2)$$

2.2 Probabilistic Graphical Model of MRI

Quantitative analysis of tissue properties with MRI is hindered by spatially correlated nonlinearity and varying noise properties of the imaging system. A parametric model that captures the variability of the imaging system, often associated with prior probability distributions of the variables of interest (and eventually of the parameters) is common practice for automated classification of MR images [9]. In this context, models of the imaging system based on finite mixtures assume that the MR image intensity is the uncertain expression, described by a parametric function, of a hidden variable that takes values in a discrete set. Since Rician noise introduced by the MRI imaging system is well approximated, for high SNR, by a Gaussian distribution [9], a Gaussian Finite Mixture model is commonly adopted to represent the MR imaging system. Spatially correlated nonlinearity of the image due to non-uniformity of the magnetic field is not taken into account in the following, though the mixture model may be extended to account for nonlinearity as in [9]. The hidden state in voxel $b \in \{1, 2, \dots, N_b\}$ is denoted by $k_b = k$, with $k \in \{1, 2, \dots, N_k\}$. MRI intensity in voxel b is denoted by $x_b = x$, with $x \in \mathbb{R}^+$.

$$p(x_b|k_b) = \mathcal{N}(x_b, \mu_{x_k}, \sigma_{x_k}) \quad (3)$$

Assuming that the prior probability of k is a multinomial distribution $p(k) = \pi_k$ and regarding the unknown parameters as random variables, the mixture model that describes MRI image formation is represented by the Directed Acyclical Graph in figure 1-B. For image segmentation the parameters of this model are commonly estimated by the EM algorithm [9].

2.3 Probabilistic Graphical Model of SPECT/MRI

Since activity is not directly observable, it is not possible to define empirically a model that expresses the interdependence of the two imaging modalities. We postulate that there is an underlying variable that, if known, renders the two

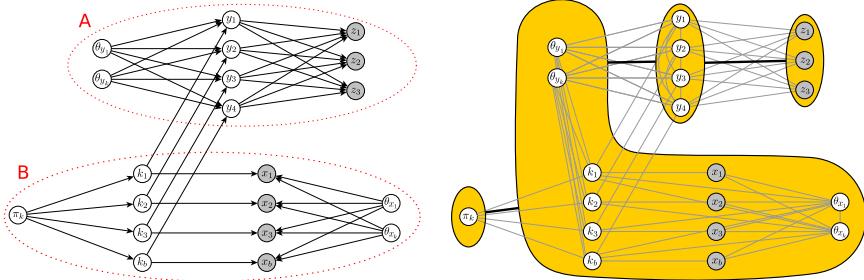


Fig. 1. Directed acyclical graph (left) and moralized Markov Network (right) of the joint generative model for the SPECT/MRI imaging system. Observed quantities are shaded. A hidden anatomical/functional state k_b at each voxel determines activity y and MRI image intensity x ; the probability distributions of x and y are independent if the hidden state $k_b = k$ is known and their probability distribution is a parametric function of parameters θ .

images independent. Indexing with $k = \{1, 2, \dots, N_k\}$ the value of a discrete hidden state, it is assumed that the hidden state in one voxel determines relaxation time and pharmaceutical concentration in the same voxel, which implies that the two variables are independent conditionally to the hidden state $k_b = k$ in b . Accounting for complexity of the reasons of variability of relaxation time and activity in a region defined by a given state, their probability distribution is assumed to be a Gaussian of unknown parameters. The conditional independence $x \leftarrow k \rightarrow y$ implies that $p(x, y)$ is a bivariate Gaussian mixture (GM) with diagonal covariance matrix, represented by the DAG in figure 1-left, which moralizes to the Markov Network on the right. In MRI imaging the intrinsic contrast is produced by differences in proton density and MR relaxation times. However, by selection of appropriate magnetization schemes, signal intensity can be modulated by other important processes such as tissue perfusion, Brownian water motion, tissue oxygenation, mass of molecular groups [8]. The distribution of image intensity is related to the hidden variables that characterized the underlying processes and regions of the image that are distinguishable from others correspond to discrete states of the underlying variable. In the proposed joint model the hidden state captures the relation between activity and MRI intensity and it assumes different meanings depending on the MRI sequence that is adopted. Examples of hidden states are normal tissue with water content corresponding to gray matter, normal tissue with water content of white matter, hypoactive gray matter tissue.

2.4 Inference

The parameters of the model that maximize the joint *pdf* correspond to the state of the system such that the observed quantities (MRI image and photon counts) are more likely to be observed. Factorization of the joint probability distribution according to the moralized graph in figure 1-right allows us to develop

an iterative method to maximize the joint pdf. Besag introduced the Iterated Conditional Modes (ICM) iterative algorithm to optimize the joint probability of a Markov Network and showed that it provides estimates with always increasing joint probability, thus converging to a local maximum. ICM consists in finding a new estimate of the unknown variable at a node of a Markov Network by maximizing its probability conditional to the neighboring nodes, given their provisional estimates. When applied to each node in turn, moving along the neighboring structure, this procedure defines a single cycle of an iterative algorithm for estimation of all the variables. Considering the factorization in figure 1-right (yellow), ICM consists in finding alternately the parameters of the GM which have highest probability given the activity (i), and the activity with highest probability given the parameters and the SPECT projection data (ii).

(i) Given activity, the DAG in figure 1, represents a bivariate GM with diagonal covariance matrix (because of the condition $y \perp x|k$). We adopt EM to compute a new estimate of the parameters that increases the jpdf as it has a well known formulation for mixture problems [9]: Denoting by π_k the prior probability of k , by μ_{x_k}, μ_{y_k} the expected values and $\sigma_{x_k}, \sigma_{y_k}$ the variances of each class k :

$$\hat{p}_{bk}^{(n+1)} = p(k_b|x_b, y_b) = \frac{\mathcal{N}(x_b, \hat{\mu}_{x_k}^{(n)}, \hat{\sigma}_{x_k}^{(n)})\mathcal{N}(y_b, \hat{\mu}_{y_k}^{(n)}, \hat{\sigma}_{y_k}^{(n)})\hat{\pi}_k^{(n)}}{\sum_{k=1}^{N_k} \mathcal{N}(x_b, \hat{\mu}_{x_k}^{(n)}, \hat{\sigma}_{x_k}^{(n)})\mathcal{N}(y_b, \hat{\mu}_{y_k}^{(n)}, \hat{\sigma}_{y_k}^{(n)})\hat{\pi}_k^{(n)}} \quad (4)$$

$$\hat{\mu}_{x_k}^{(n+1)} = \frac{1}{N_b} \frac{\sum_{b=1}^{N_b} \hat{p}_{bk}^{(n+1)} x_b}{\hat{\pi}_k^{(n+1)}} \quad \hat{\sigma}_{x_k}^{2(n+1)} = \frac{1}{N_b} \frac{\sum_{b=1}^{N_b} \hat{p}_{bk}^{(n+1)} (\hat{\mu}_{x_k}^{(n+1)} - x_b)^2}{\hat{\pi}_k^{(n+1)}} \quad (5)$$

$$\hat{\mu}_{y_k}^{(n+1)} = \frac{1}{N_b} \frac{\sum_{b=1}^{N_b} \hat{p}_{bk}^{(n+1)} y_b}{\hat{\pi}_k^{(n+1)}} \quad \hat{\sigma}_{y_k}^{2(n+1)} = \frac{1}{N_b} \frac{\sum_{b=1}^{N_b} \hat{p}_{bk}^{(n+1)} (\hat{\mu}_{y_k}^{(n+1)} - y_b)^2}{\hat{\pi}_k^{(n+1)}} \quad (6)$$

$$\hat{\pi}_k^{(n+1)} = \frac{1}{N_b} \sum_{b=1}^{N_b} \hat{p}_{bk}^{(n+1)} \quad (7)$$

(ii) Given the parameters of the GM, the probability of activity is the product of two terms (figure 1): $p(y|x, \theta, \pi, z) = p(y|x, \theta, \pi)p(z|y)$ where $p(z|y)$ is the Poisson likelihood of equation (2). This is maximized by the One Step Late (OSL) EM algorithm introduced by Green [11]:

$$\hat{y}_b^{(n+1)} = \hat{y}_b^{(n)} \frac{1}{\sum_{d=1}^{N_d} p_{bd} + \frac{\partial}{\partial y_b} \log p(y|x, \theta, \pi) \Big|_{y_b^{(n)}}} \sum_{d=1}^{N_d} \frac{p_{bd} z_d}{\sum_{b'=1}^{N_b} p_{b'd} \hat{y}_{b'}^{(n)}} \quad (8)$$

$p(y|x, \theta, \pi)$ is obtained by marginalizing over k :

$$p(y|x, \theta, \pi) = \prod_{b=1}^{N_b} p(y_b|\theta, x) = \prod_{b=1}^{N_b} \sum_{k=1}^{N_k} \pi_k p(y_b|k_b, \theta, x) = \prod_{b=1}^{N_b} \sum_{k=1}^{N_k} \pi_k \mathcal{N}(y_b, \mu_{y_k}, \sigma_{y_k})$$

By the chain rule of differentiation, the gradient in (8) simplifies to:

$$\frac{\partial}{\partial y_b} \log p(y|\theta_y) = \sum_{k=1}^{N_k} \pi_k \frac{\mu_{y_k} - y_b}{\sigma_{y_k}^2} p_{bk} \quad (9)$$

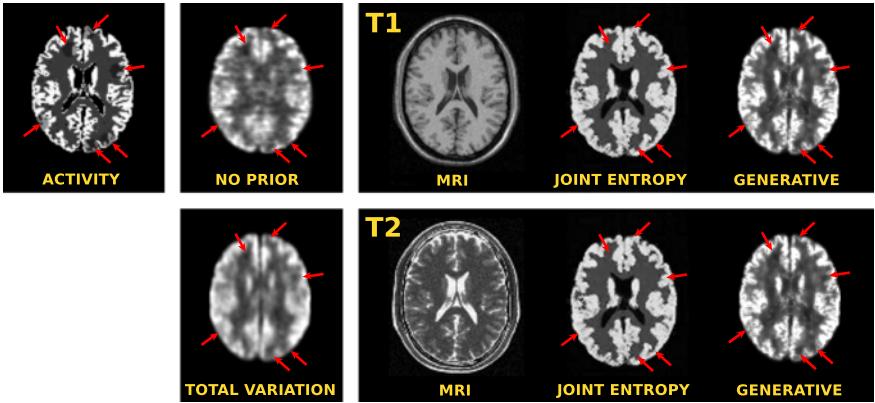


Fig. 2. Central transaxial slice of the activity phantom, the T1 and T2-weighted MRI images and reconstructed activity. The red arrows point to the simulated cold-spots.

3 Validation Study

Synthetic brain data from the BrainWeb [12] database was adopted in order to validate the proposed reconstruction algorithm and compare it with other methods. The MRI and functional imaging processes were decoupled by adopting the normal brain tissue model from the database as ground truth of tissue composition. T1 and T2-weighted MRI images were generated with the BrainWeb simulator, which realistically accounts for noise of the imaging system. The parameters of the simulator were set for noise standard deviation at 3% of the brightest tissue and perfect uniformity of the magnetic field (in accordance with the simplistic GM model). Brain perfusion was simulated by associating typical activity levels to different tissue types, proportionally to partial voxel occupation. Specifically the activity in gray matter was set to a value 4 times higher than in all other tissues. The total number of counts was set to 2.5 Million. 5 spherical cold-spots (red arrows in figure 2) of equal size were simulated at random locations centered on the central transaxial plane by lowering the activity by 30%. The SPECT imaging system was simulated by means of a rotation-based projector with realistic Collimator-Detector Response (CDR) and applying Poisson noise to the projections. The parameters of the imaging system were set to emulate a SPECT imaging system based on GE Infinia with Low Energy High Resolution (LEHR) collimator (size of the detector plane: $540 \times 400 \text{ mm}$; point spread function full width at half maximum (FWHM): $\text{FWHM}@20\text{mm} = 5.11 \text{ mm}$, $\text{FWHM}@160\text{mm} = 9.98 \text{ mm}$; distance of the detector from the axis of rotation: 133 mm ; 120 positions of the gamma camera from 0 to $2\pi \text{ rad}$). The MRI and activity images were defined on a cubic grid of $(128 \times 128 \times 128)$ voxels.

The number of tissue types was assumed to be $N_k = 4$; μ_{x_k} were initialized to evenly spaced values in the range of intensity of the MRI image; σ_{x_k} were initialized to $1/N_k$ of the image intensity range; μ_{y_k} were initialized to evenly spaced

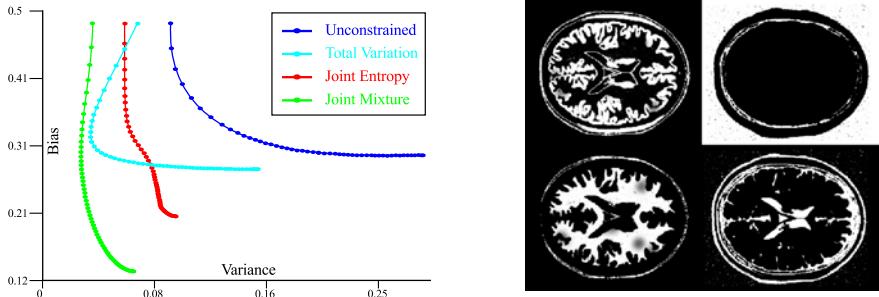


Fig. 3. Left: *bias/variance* at each iteration step for SPECT combined with T1-weighted MRI. Right: Anatomical/functional classification $p(k_b|x_b, y_b)$ for SPECT combined with T1-weighted MRI.

values between 0 and the maximum activity assigned to the phantom; σ_{y_k} to $1/N_k$ of the maximum activity assigned to the phantom; the mixing coefficients to $\pi_k = 1/N_k \forall k \in N_k$. $N1$ and $N2$ were set to 1. A *bias/variance* characterization of the reconstruction algorithms was performed by generating multiple instances of the sinogram data (figure 3-right). The algorithm was launched on 15 realizations of the sinogram. Ensemble *bias* and *variance* were calculated according to the definition in [13].

For comparison, activity was estimated with unconstrained MLEM, with a prior based on total variation (TV) and with a prior based on Joint Entropy (JE) [1]. The hyper-parameters of the TV and JE priors were chosen in order to optimize *bias/variance* as described in [13]. The curves in figure 3-right report *bias/variance* for 100 iterations of each of the methods. The proposed algorithm outperforms all the aforementioned methods in terms of *bias/variance* and identification of lesions by visual assessment. The estimates of the parameters after convergence were inspected and μ_{x_k} and μ_{y_k} demonstrated to converge to large areas of normal activity. None of the classes converged to the cold-spot values. While this issue needs to be addressed by specific solutions, the cold-lesions still appear more visible than with the other reconstruction algorithms. This means that information from the photon counts keeps activity in the cold-spots low, though the anatomical side of the model would tend to suppress them. If the lesions are more visible it is due to overall better redistribution of activity.

4 Discussion

We have introduced a unified framework based on a probabilistic joint generative model of a combined SPECT/MRI imaging system and we have described an iterative algorithm to estimate the parameters of the model, producing an estimate of activity that accounts for prior information from the MRI image along with multi-modal tissue classification. The proposed model is based on the assumption that activity and relaxation time are related because of the existence of a finite number

of states; the phantom has been generated accordingly assigning uniform activity to each of the ground-truth tissue types. The improvement of *bias/variance* and lesion identification over other methods demonstrate that the method works well when the assumption that the model relies on is verified. It is difficult, because of the lack of real life integrated multi-modal imaging systems and of empirical models of the interaction of the pharmaceutical with MRI related tissue properties, to validate the method under realistic conditions. PGM's provide a powerful formalism for model definition and inference in multi-modal imaging systems and offer the potential to integrate estimation at image level with more complex decisioning systems. Extensions of the model proposed in this paper include estimation (correction) of non-linear response of the MRI imaging system, spatial dependence assumption of k and the use of priors for the parameters.

References

1. Atre, A., Vunckx, K., Baete, K., Reilhac, A., Nuyts, J.: Evaluation of different MRI-based anatomical priors for PET brain imaging. In: IEEE Nucl. Sci. Sym. Conf., Orlando, pp. 1–7 (October 2009)
2. Leahy, R., Yan, X.: Incorporation of Anatomical MR Data for Improved Functional Imaging with PET. In: Inf. Proc. in Med. Imag., pp. 105–120. Springer, Heidelberg (1991)
3. Scheines, R.: An Introduction to Causal Inference. In: McKim, V., Turner, S. (eds.) Causality in Crisis?, pp. 185–200. University of Notre Dame Press
4. Venkataraman, A., Rathi, Y., Kubicki, M., Westin, C.-F., Golland, P.: Joint Generative Model for fMRI/DWI and Its Application to Population Studies. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010 Part I. LNCS, vol. 6361, pp. 191–199. Springer, Heidelberg (2010)
5. Hiltunen, P., Prince, S.J.D., Arridge, S.: A combined reconstruction-classification method for diffuse optical tomography. Phys. Med. and Biol. 54, 6457–6476 (2009)
6. Sastry, S., Carson, R.E.: Multimodality Bayesian algorithm for image reconstruction in positron emission tomography: a tissue composition model. IEEE Trans. on Med. Imag. 16(6), 750–761 (1997)
7. Rangarajan, A., Hsiao, I.T., Gindi, G.: A Bayesian joint mixture framework for the integration of anatomical information in functional image reconstruction. J. of Math. Imag. and Vis. 12(3), 199–217 (2000)
8. Blamire, A.M.: The technology of MRI - the next 10 years? The British J. of Radiology 81, 601–617 (2008)
9. Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P.: Automated model-based tissue classification of MR images of the brain. IEEE Trans. on Med. Imag. 18(10), 897–908 (1999)
10. Ashburner, J., Friston, K.J.: Unified segmentation. Neuroimage 26(3), 839–851 (2005)
11. Green, P.G.: Bayesian Reconstructions From Emission Tomography Data Using a Modified EM Algorithm. IEEE Trans. on Med. Imag. 9(1), 84–93 (1990)
12. Brain Web, <http://mouldy.bic.mni.mcgill.ca/brainweb/>
13. Pedemonte, S., Cardoso, M.J., Bousse, A., Panagiotou, C., Kazantsev, D., Arridge, S., Hutton, B.F., Ourselin, S.: Class conditional entropic prior for MRI enhanced SPECT reconstruction. In: IEEE Nucl. Sci. Sym. Conf., Knoxville, pp. 3292–3300 (November 2010)

Directed Graph Based Image Registration

Hongjun Jia¹, Guorong Wu¹, Qian Wang^{1,2}, Yaping Wang^{1,3}, Minjeong Kim¹,
and Dinggang Shen¹

¹ Department of Radiology and BRIC, University of North Carolina at Chapel Hill
`{jiahj, grwu, mjkim, dgshen}@med.unc.edu`

² Department of Computer Science, University of North Carolina at Chapel Hill
`qianwang@cs.unc.edu`

³ Department of Automation, Northwestern Polytechnical University, Xi'an, China
`ypwang@email.unc.edu`

Abstract. In this paper, a novel intermediate templates guided image registration algorithm is proposed to achieve accurate registration results with a more appropriate strategy for intermediate template selection. We first demonstrate that registration directions and paths play a key role in the intermediate template guided registration methods. In light of this, a directed graph is built based on the asymmetric distances defined on all ordered image-pairs in the dataset. The allocated directed path can be used to guide the pairwise registration by successively registering the underlying subject towards the template through all intermediate templates on the path. Moreover, for the groupwise registration, a minimum spanning arborescence (MSA) is built with both the template (the root) and the directed paths (from all images to the template) determined simultaneously. Experiments on synthetic and real datasets show that our method can achieve more accurate registration results than both the traditional pairwise registration and the undirected graph based registration methods.

1 Introduction

Non-rigid registration between images with large deformation is challenging due to the difficulty in establishing the precise and consistent correspondences and optimizing the highly non-linear energy function [1, 2]. Recently, several methods have been proposed to alleviate this difficulty by getting help from other images [3, 4]. For example, to warp subject S to template T in a dataset \mathcal{I} , a set of intermediate templates (IT), $\mathcal{T} = \{T_1, \dots, T_m\} \subseteq \mathcal{I}$, can be selected to guide the registration from S to T , *i.e.*, $S \rightarrow T_1 \rightarrow \dots \rightarrow T_m \rightarrow T$. The deformation between S and T is thus decomposed into a series of small ones, and the subject will travel through the path by accumulating these gentle deformations and finally arrive at the template space. Since the difference between neighboring images on the path is smaller than that between S and T , the registration from S to T along the path takes less risk of being trapped in local minima than their direct registration. Here, the registration from S to T is denoted by $\mathcal{R}_{S \rightarrow T}$.

It is worth noting that \mathcal{T} is essentially an ordered set and the registration from S to T is also a directed process. That means only forward registrations (*e.g.*, $\mathcal{R}_{S \rightarrow T_1}$, $\mathcal{R}_{T_i \rightarrow T_{i+1}}$, and $\mathcal{R}_{T_m \rightarrow T}$, $i = 1, \dots, m-1$) are directly related to the final result, while the

registrations between other image pairs and backward registrations (*e.g.*, $\mathcal{R}_{T_1 \rightarrow S}$) are irrelevant. Moreover, the measures to evaluate the performance on $\mathcal{R}_{S \rightarrow T}$ are generally different from that of $\mathcal{R}_{T \rightarrow S}$, and sometimes this difference is significant. Different selections of intermediate templates in T can also affect the registration accuracy greatly as we can see in the experiments below. Based on these observations, we introduce a new term, *directionality*, to describe the effect related to the registration paths and directions, which is proven to be a key factor and can facilitate the selection of more appropriate intermediate templates to better guide the registration.

The directionality exists in *not only* pairwise *but also* groupwise registration. For example, a graph based method is applied in [3, 4] to build a tree on the dataset, and each registration from one node to its neighbor is a directional warping procedure. To be consistent with the registration step, the directionality should also be taken into consideration in the tree-building step, *e.g.*, constructing a directed graph with different weights assigned to the edges from S to T and from T to S . However, the metrics defined in [3, 4] are symmetric by averaging two directional registration results for building an undirected graph. Such non-negligible inconsistency may undermine the accuracy and robustness of the final registration results, since the directed registration step requires a corresponding directional-intermediate-templates-selection step during path formulation. Some previous works studied the effect of the order of moving subject and fixed template in the pairwise registration, *e.g.*, the consistent image registration [1] that estimates an inverse consistent deformation field, the symmetric diffeomorphic registration [5] that warps two images towards an implicit space in-between and the asymmetric image-template registration [6] that introduces a correction factor to the symmetric cost function. However, the directionality on image population has not been fully studied in the literature to our best knowledge.

In this paper, a novel directed graph based image registration method is proposed. First, we define an asymmetric similarity measure on ordered image pairs and construct a fully connected directed graph. To learn the local data structure, a directed kNN graph is then extracted and the optimal paths from one image to another are allocated. The pairwise registration can be done by registering the subject to the template along the allocated path. For the groupwise registration on an image population, the minimum spanning arborescence (MSA) [7] is built with the root node (template) determined automatically. Experimental results on synthetic dataset and real brain MR image datasets demonstrate that the proposed algorithm can achieve more consistent and robust registration results than other widely applied methods.

2 Method

In this section, the proposed directed graph based registration framework is detailed. We first explore the directionality in image registration and then apply it in both IT-guided pairwise and groupwise registrations.

2.1 The Importance of Directionality in Registration

To register two images S and T together, either of them can serve as the fixed template and the other one as the moving subject. However, due to the non-existence of a

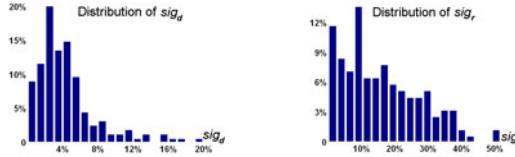


Fig. 1. The distribution of the directionality significance defined on the intensity difference between the registered image and the template (sig_d , left) and the smoothness of the estimated deformation field (sig_r , right) on a set of 18 real brain images.

perfect non-rigid registration solution, the registration performance of forward and backward registrations could be different, which is especially true when two images have very different anatomical structures.

To further quantitatively analyze this difference, we first define a directed distance on the ordered image pair (S, T) based on the result of $\mathcal{R}_{S \rightarrow T}$. The optimal deformation field $f_{S \rightarrow T}(x)$ is obtained by maximizing the similarity between T and the warped warped image $S(f_{S \rightarrow T})$. Two directed measures are defined on $\mathcal{R}_{S \rightarrow T}$: the intensity difference, $d(S(f_{S \rightarrow T}), T) = \int_{\Phi} |S(f_{S \rightarrow T}(x)) - T(x)|^2 dx$, and the smoothness of the deformation field: $r(f_{S \rightarrow T}) = \int_{\Phi} \|\nabla^2 f_{S \rightarrow T}(x)\| dx$, where Φ is the image domain and ∇^2 is the Laplacian operator. The final directed difference can be calculated as a weighted average with $\alpha \in [0, 1]$, i.e.,

$$g(S, T) \triangleq \alpha \cdot d(S(f_{S \rightarrow T}), T) + (1 - \alpha) \cdot r(f_{S \rightarrow T}). \quad (1)$$

The two measurements are normalized into the same range before being used in Eq. 1. The directed difference g defines an asymmetric measurement between two images, and it is different from a true metric which has to be symmetric. To quantitatively measure the directionality of a directed metric, we define the relative difference between two directed measurements as the directionality significance. For the measurement d as defined above, its directionality significance sig_d on the image pair (S, T) (S, T) is,

$$sig_d(S, T) = \frac{|d(S(f_{S \rightarrow T}), T) - d(T(f_{T \rightarrow S}), S)|}{\min(d(S(f_{S \rightarrow T}), T), d(T(f_{T \rightarrow S}), S))} \times 100\%. \quad (2)$$

We can also define sig_r in a similar way. For illustration, the directionality significances sig_d and sig_r are calculated for all image pairs in a dataset of 18 elderly brains with large shape difference. The distribution of the directionality significance is shown in Fig. 1. We can see that the relative difference between the measurement on the forward and backward registration results can be up to 20% and 50% on sig_d and sig_r , respectively, implying that the ignorance of the directionality in image registration may lead to a sub-optimal solution.

2.2 Directed Graph Based Pairwise Registration

Given two subjects I_s and I_t in an image set $I = \{I_i | i = 1, 2, \dots, N\}$, $1 \leq s, t \leq N$, the goal of IT-guided pairwise registration is to warp I_s to I_t with the help of a selected

subset of images in \mathcal{I} . After assigning the directed dissimilarity g to each ordered image pair (I_s, I_t) , we can build a directed graph by considering each subject as a node and weighting each directed edge $e_{s \rightarrow t}$ with $g(I_s, I_t)$ as defined in Eq.1. To further learn the underlying structure, we construct a directed kNN graph and then approximate the directed distance on the image space by the shortest directed path from I_s to I_t on the kNN graph. The construction of directed kNN graph is similar to that of undirected kNN graph. The difference is that, in the directed kNN graph, two types of nearest neighbors are considered: 1) for each I_i , we first sort all I_u ($u \neq i, 1 \leq u \leq N$) by $g(I_i, I_u)$, then the nearest k neighbors with the shortest distances are selected as k-out-NN (the edge direction is from I_i to I_u); 2) another set of k-in-NN (from I_u to I_i) is also selected. For each I_i , only those directed edges from k-in-NN to the subject I_i and those from the subject I_i to k-out-NN are kept in the directed kNN graph.

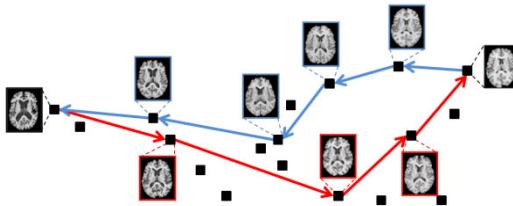


Fig. 2. Different paths determined by directed graph based method for the forward and backward registrations are shown. Only the images on the path are displayed.

Next, we adopt the Floyd-Warshall algorithm to find the shortest paths between any ordered image pair on the directed kNN graph. For a given image pair (I_s, I_t) , the registration can be implemented by following the determined directed registration path on the image space. It should be noted that the shortest directed path from I_s to I_t may be different from that in opposite direction, due to the characteristics of the directionality. For example, with an 18-elderly-brain dataset, we have detected totally 153 pairs of directed paths. Among them, only 40 pairs (26.1%) share exactly the same intermediate templates with each other, and all others (113 pairs, 73.9%) have different sets of intermediate templates to guide the respective directed registration. This shows the existence of the directionality in the pairwise registration problem of real images. Such an example on the elderly brain image set is given in Fig. 2, which shows the different directed registration paths allocated between two images.

2.3 MSA Based Groupwise Registration

To solve the groupwise registration in a population, both the final template and the intermediate templates for each subject need to be determined. With the assigned asymmetric weights on the directed edges, the MST algorithm is no longer applicable since it is only defined on the undirected graph. Instead, we propose to build an MSA on the dataset. Note that in other tree-based groupwise registrations, the global template selection and respective registration paths allocation are two separate steps based on different criteria by focusing on either the importance of the final template representativeness or the overall registration error. We build MSA to solve the

combinative optimization problem by simultaneously determining the best template and the optimal path to the template for each subject.

In graph theory, the MSA algorithm is first proposed by Zhu and Liu [7] to minimize the total length of the directed edges from all nodes to the root. The original MSA method can only construct a tree with a pre-specified root with $O(N^2)$ complexity. However, if the template is predefined to guide the path allocation, we can only obtain a sub-optimal solution. To pursuit the global optimal solution of MSA, we have to set each of the subjects as the potential root and build N trees totally, and then choose the tree giving the shortest total path length as the final solution. Thus, the computational complexity increases to $O(N^3)$. Here, we take another strategy to find the global optimal solution in one run by adding a virtual node, which is designated as the fixed root in the extended graph. N extra directed edges from all existing N nodes to the virtual node are added with the same weight, which is assigned to be larger than the summation of all edge lengths in the original graph. An MSA is built on the extended graph with a fixed root (the virtual node). Then, the global optimal MSA on the original graph can be extracted by removing the virtual node from the extended tree. The reason to account for the mathematical equivalence between these two methods is that all directed edges connecting the virtual node have the same weight, and there is one and only one such edge can be selected in the extended tree due to the minimum total length requirement (see the supplementary file). The computational complexity of this new solution is $O((N + 1)^2)$, i.e., at the same level as in the case of building MSA with a predefined root node.

2.4 Discussion

The proposed directed graph based registration has two major advantages compared with other methods. First, the IT selection step and the registration step are inherently *consistent* by considering the directionality. Second, the template in groupwise registration is determined together with the registration paths, which means our method can take both the final template's representativeness and the overall registration error into consideration within a single framework. Note that the computational complexity of the proposed algorithm is same as other tree-based algorithms. We need $O(N^2)$ times of fast registrations for the graph construction and $O(N)$ times of elaborated registrations for the final registration by sequentially compositing the deformation fields at each node along the path. In our experiments, the directed kNN graph built on the results of the elaborated registration is quite similar to that based on the fast registration, but the computing time is about 3~4 times longer.

3 Experiments

To demonstrate the advantage of the proposed directed graph based registration method, we evaluate it on both synthetic dataset and real brain MR images together with two highly related methods, *i.e.*, the undirected graph based registration [3, 4] and the traditional pairwise registration, and two widely used groupwise registration methods, *i.e.*, the group mean method [8] and the congealing method [9]. Throughout this paper, the diffeomorphic demons [2] is used as the basic pairwise registration tool.

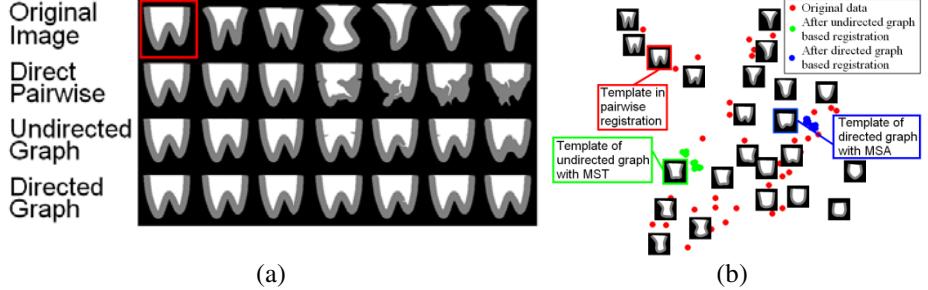


Fig. 3. Registration results on a synthetic dataset by different methods to a pre-defined template as shown in the red box (a). The data distribution on the 2D PCA space is illustrated in (b).

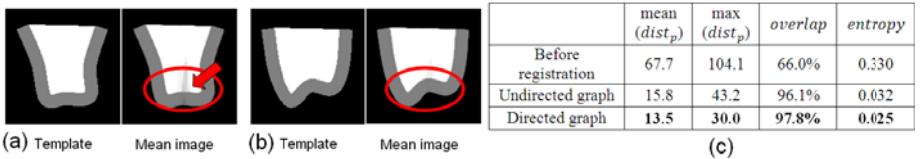


Fig. 4. The mean images after groupwise registration by the undirected (a) and directed (b) graph based methods are shown together with their respective template. The quantitative comparisons on average distance, overlap rate, and entropy are also provided in (c).

3.1 Synthetic Dataset

We first test on a synthetic dataset with 56 2D shapes simulating gyri and sulci with samples shown in Fig. 3a. The top-left image is pre-selected as the template in pairwise registration to evaluate the performance on registering images with large differences. As shown in Fig. 3a, the registration results vary dramatically by following different registration paths. Several images even fail to be registered to the template by the traditional pairwise registration. On the contrary, the IT-guided registration can provide much better results, although the undirected graph based method are not very satisfactory in some cases when the shape difference is considerably large. The results in this experiment clearly show that even both with the guidance, the directed graph base method can find more suitable paths than its undirected counterpart.

The influence of the directionality is also validated in IT-guided groupwise registration. The root in MST on the undirected graph is selected to be the node with the shortest total length to all other nodes. Here we choose the same weighting factor $\alpha = 0.5$ in Eq. 1. Each original image is projected onto a 2D PCA space as illustrated in Fig. 3b, and we can see that the root template of MSA is selected to be much closer to the population center than that of MST. The registered images of different methods are also projected onto the same 2D space to visualize the registration results. It can be seen that the registration results of directed graph based method are much denser around the geometric center of the underlying space.

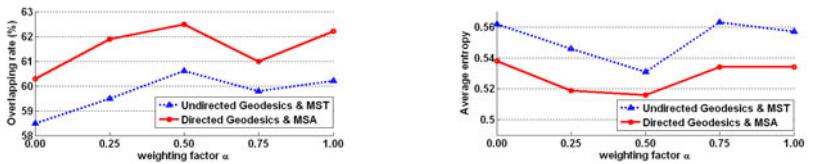


Fig. 5. The average overlap rate (left) and average entropy (right) on ADNI dataset with different weighting factors $\alpha=0.0, 0.25, 0.5, 0.75$, and 1.0

The mean aligned images by two methods are shown together with the respective template in Fig. 4a and 4b. The undirected method fails to register some images precisely thus giving a mean image with a much larger shaded area in the center (as indicated by the arrow). The proposed method can achieve more consistent registration. We also provide the quantitative comparisons in Fig.4c on the mean/max of all the pairwise distances among registered images ($dist_p$), the average overlap rate on the common space measured by Jaccard coefficient, and the average entropy on the region within the red circle in Fig. 4a and 4b. We can see that the proposed method consistently outperforms the undirected graph based method. The paired t -test on the pairwise distances between registered images also shows that the directed graph based method can generate a much more compact registered image set (with $p < 10^{-3}$).

3.2 Real Brain MR Image Datasets

We apply the proposed algorithm on ADNI dataset to further test its performance on a large dataset with real images. 60 brain MR images are selected with 20 from each category of normal control, MCI and AD. To show the influence of weighting factor in Eq. 1, both directed and undirected graph based methods are tested with $\alpha = 0.0, 0.25, 0.5, 0.75$ and 1.0 , respectively. From Fig. 5, we can see that the registration accuracy measured by the average overlap rate and entropy on all registered tissue-segmented images by the proposed method is consistently better than the undirected graph based method. Since both methods select the same template with $\alpha = 0.5$, we can fairly compare these two methods in this case. The quantitative comparison on tissue overlap rate and average entropy clearly shows the advantage of the proposed method in determining the registration paths. Our method gives 62.5% on the average tissue overlap rate, which is much higher than 60.5% given by the undirected graph based registration and 57.5% by the traditional pairwise registration to the same template. This result is also significantly higher than results of the group mean method [8] (58.7%) and the congealing method [9] (47.2%). We also measure the overlap rates on different tissues, GM, WM and CSF, on ADNI dataset in Table 1, and our method achieved the top result on all three tissues.

We also apply the proposed algorithm on LONI LPBA40 dataset including 40 brain images with 54 labeled ROIs. We set $\alpha = 0.5$ in Eq. 1 for all tests. Our method can increase the mean overlap rate on the aligned images to 68.2%, higher than the MST based method (66.3%). Moreover, our method is better for most ROIs (48 out of all 54 ROIs), and the paired t -test shows that it can achieve consistently better alignment than undirected graph based method (with $p < 0.001$).

Table 1. The overlap rates and average entropy of the registered segmentation images by five different methods.

	Overlap rate (%)			Entropy
	GM	WM	CSF	
Before registration	40.1	54.5	28.8	0.85
Direct pairwise	54.7	70.4	47.4	0.58
Undirected graph	57.3	73.2	50.8	0.54
Directed graph	59.6	74.8	52.9	0.52
Group mean	54.1	73.4	48.5	0.55
Congealing	42.9	59.5	39.2	0.59

To fairly compare the performance of different registration methods, we use the same template to all three methods, *i.e.*, the traditional pairwise registration, the undirected graph based registration, and the directed graph based registration. Specifically, the template determined by MSA is used as the common template, and the average overlap rate given by the direct pairwise registration and the undirected graph based method is 64.8% and 66.7%, respectively. Our method can still achieve the best result (68.2%) as listed in Table 2. One possible reason is that, for the same template, MSA can find better registration paths than the undirected MST based solutions with the directionality being considered. We can also make comparison between the results by undirected graph based registration using different templates, to demonstrate the importance of selecting an appropriate template in registration. Even with the same undirected graph based tree building algorithm, the performance of the registration with the template selected by MSA is slightly better than that of the registration with the template selected by MST. This also implies the better representativeness of the template selected by the directed graph based method.

Table 2. Comparison of different methods and the effect of template on LPBA40 dataset. T_{MSA} and T_{MST} are the templates selected by MSA and MST, respectively.

	Path Allocation	Template	Overlap	Entropy
Direct Pairwise	Direct path	T_{MSA}	64.8%	0.462
Undirected Graph	MST	T_{MST}	66.3%	0.448
Undirected Graph (*)	MST	T_{MSA}	66.7%	0.445
Directed Graph	MSA	T_{MSA}	68.2%	0.431
Group Mean	-	-	66.3%	0.450
Congealing	-	-	66.8%	0.449

4 Conclusion

A new image registration framework is proposed by taking advantages of the directionality in image registration. The pairwise registration is guided along the shortest path on the directed graph, and the groupwise registration is achieved by building an MSA with the automatically selected template. The proposed algorithm can determine more appropriate intermediate templates to facilitate more accurate registration. Our future work includes applying our method to the general groupwise registration of large clinical dataset for detecting disease related brain abnormalities.

References

1. Christensen, G.E., Johnson, H.J.: Consistent Image Registration. *IEEE Transactions on Medical Imaging* 20, 568–582 (2001)
2. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: efficient non-parametric image registration. *Neuroimage* 45, S61–S72 (2009)
3. Munsell, B.C., Temlyakov, A., Wang, S.: Fast Multiple Shape Correspondence by Pre-Organizing Shape Instances. In: *IEEE Conference on CVPR*, pp. 840–847 (2009)
4. Hamm, J., Ye, D.H., Verma, R., Davatzikos, C.: GRAM: A framework for geodesic registration on anatomical manifolds. *Medical Image Analysis* 14, 633–642 (2010)
5. Avants, B., Gee, J.C.: Geodesic estimation for large deformation anatomical shape averaging and interpolation. *NeuroImage* 23, S139–S150 (2004)
6. Sabuncu, M.R., Yeo, B.T.T., Van Leemput, K., Vercauteren, T., Golland, P.: Asymmetric image-template registration. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) *MICCAI 2009. LNCS*, vol. 5761, pp. 565–573. Springer, Heidelberg (2009)
7. Zhu, Y.J., Liu, T.H.: On the shortest arborescence of a directed graph. *Science Sinica* 14, 1396–1400 (1965)
8. Joshi, S., Davis, B., Jomier, M., Gerig, G.: Unbiased Diffeomorphic Atlas Construction for Computational Anatomy. *NeuroImage* 23, S151–S160 (2004)
9. Learned-Miller, E.G.: Data Driven Image Models through Continuous Joint Alignment. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28, 236–250 (2006)

Improving the Classification Accuracy of the Classic RF Method by Intelligent Feature Selection and Weighted Voting of Trees with Application to Medical Image Segmentation

Mohammad Yaqub^{1,2}, M. Kassim Javaid², Cyrus Cooper², and J. Alison Noble¹

¹ Institute of Biomedical Engineering, Dept. of Engineering Science, University of Oxford

² Nuffield Dept. of Orthopaedics, Rheumatology & Musculoskeletal Sciences,
University of Oxford

Abstract. Enhancement of the Random Forests to segment 3D objects in different 3D medical imaging modalities. More accurate voxel classification is achieved by intelligently selecting "good" features and neglecting irrelevant ones; this also leads to a faster training. Moreover, weighting each tree in the forest is proposed to provide an unbiased and more accurate probabilistic decision during the testing stage. Validation is performed on adult brain MRI and 3D fetal femoral ultrasound datasets. Comparisons between the classic Random Forests and the proposed new one show significant improvement on segmentation accuracy. We also compare our work with other techniques to show its applicability.

Keywords: Random forests, machine learning, feature selection, brain MRI segmentation, 3D fetal ultrasound segmentation.

1 Introduction

Random Forests (RF) is an ensemble of multiple decision trees that are trained in a random manner. We used RF as a classifier in this work although it can be used as a regressor [1]. Although the RF technique is computationally efficient, inherently multi-class, able to handle huge feature space efficiently, and easily parallelizable, it has not been widely used for medical image segmentation. To our knowledge, only a handful of conference papers have been published that use the classic technique to do medical image segmentation [2-5]. We believe there are some issues in the way that the technique works when applying it to do medical image segmentation in 3D or 4D. First, the size of typical 3D/4D medical images is large. Since RF is a discriminative classifier which relies on image features to classify objects, its success is tightly related to the discrimination power of such features. In high dimensional data, a huge number of features can be computed but a significant number of these are irrelevant. Therefore, the existence of "bad" features may drive the technique to provide a poor classification. Second, although the concept of equal opportunity (e.g., an equal vote from each tree in RF) may be highly welcomed in different decision making problems, we believe it is not the best to use in this classifier since randomness is well used during the training stage which varies the strength of the trees.

In this paper, we propose a novel segmentation technique based on the RF classifier that addresses the two above limitations and demonstrate the technique on two different medical image segmentation problems. The main contributions are 1) a solution to overcome the problem of having a huge feature space which may contain a significant number of irrelevant features and 2) a novel weighted mechanism to combine tree decisions to reduce any bias toward weaker trees.

2 Classic RF in Medical Image Segmentation and Detection

Recently, Lempitsky et al. [2] have used the standard binary RF to automatically delineate the myocardium in 3D ultrasound (US) of adult hearts. Yi et al. [3] segmented three brain tissues from 20 MRI volumes using the traditional RF technique after bias field correction. Geremia et al. [4] have used the conventional RF technique to segment multiple sclerosis in multi-channel brain MR images. In a pilot study [5], we proposed a weighted RF technique in which weights are assigned to trees during testing depending on their strength to classify a new test case. On the other hand, Criminisi et al. [1, 6] have used the standard RF technique to automatically detect several organs in CT volumes by finding a 3D bounding box around each organ. To our knowledge these are the only published technical papers that apply the classic RF to medical image segmentation and detection.

To facilitate the understanding of our novel contributions we give a short description of the RF training and testing stages. We also highlight the problems that face RF when using it to segment high dimensional data. Training an RF proceeds by building several random decision trees. Each node within a tree is a question (a classifier) and according to the answer of this question a split to the training examples happens to either the left or right. Leaf nodes contain a probabilistic outcome of answering a series of questions. RF builds several decision trees in a random nature. There are two sources of randomness in the training stage. The first is in the selection of training examples (Bagging [7]) for each tree and the second is in the feature selection to create a node in any tree. In [7, 8], Breiman discusses these two selection processes in detail. To create a node in a tree, n' features are randomly selected from the feature pool of size n . For each feature, a "good" threshold is found to form a classifier. The classifier (feature/threshold) that maximizes the expected gain of information (ΔE) when classifying the training examples V is chosen such that V becomes $(V_{\text{left}}, V_{\text{right}})$ [9, 10]. Left and right nodes are recursively trained on V_{left} , V_{right} respectively. The recursive creation of a tree stops if either 1) the maximum tree depth (L) is reached or 2) no gain of information is achieved. After training is done, T trees are created. Each decision node in any tree contains a classifier which is represented by a feature and its threshold. Each leaf node contains a probabilistic distribution of the training examples that reached it for every class.

Notice that the choice of the number of randomly selected features at each node (n') is critical. If n' is large enough then the probability to have a good feature within the n' features is high. Increasing the value n' is preferable but it is computationally expensive. Decreasing the value of n' will speed up the training process but will give lower classification accuracy because of the lower chance of choosing a good classifier at each tree node. One can argue that training time is not important

compared to testing time. This is true but the techniques developed so far to segment medical images using the standard RF use only a very small value of n' . For instance, in [4] $n'=950$ was used and roughly speaking two 3D rectangles of up to 32^3 voxels each were used as features. This gives approximately $32^6 (>10^9)$ permutations of features. This gives a low probability to have a good feature within the n' ones which leads to a poor classification accuracy. Fig. 1 shows a histogram of ΔE of Haar3D and Rectangle3D feature sets (see next section for further discussion on feature sets). From Fig. 1, we note that a small number of features are relatively "good" while many features are "bad". The performance of these features vary and a large portion of these features have low scores. In this work, we propose a more accurate selection of features during training while preserving the random nature of the technique.

Testing an unseen example (pixel/voxel) requires the test example to traverse each tree separately until a leaf node of each tree is reached. The probabilistic decision, calculated during training for the reached leaves from all trees are combined to give the final probabilistic decision. In the literature [8-10], averaging the probabilistic decisions from T trees is commonly performed to generate one probabilistic decision for each class of an unseen example v in the RF framework, see equation (1). After combining T probabilities into one, the class of the maximum probability is usually chosen to label that example [8, 10],

$$\forall c_j \in \{c_1, \dots, c_{Labels}\} \quad p(c_j | v, RF) = \frac{1}{T} \sum_{t=1}^T p(c_j | v, leaf(tree_t)). \quad (1)$$

Here $p(c_j | v, leaf(tree_t))$ is the class distribution that is estimated as a histogram of the class c_j of the training examples v that reached the leaf node ($leaf$) on the t^{th} tree.

Since each tree is built in a random manner, the decisions at leaves vary from one tree to another. Fig. 2 shows segmentation results for five individual trees in one forest. Notice that each tree provides a slightly different segmentation results depending on the classifiers inside each tree. As a result, we propose a weighted voting procedure where trees should contribute unevenly to the RF final decision.

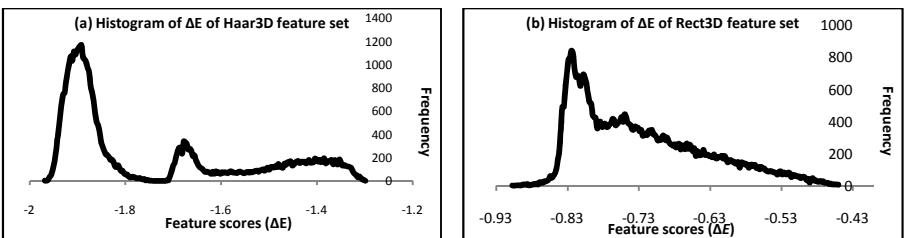


Fig. 1. Histograms of feature scores (ΔE) for two different feature sets. (a) shows a histogram of Haar3D feature set scores (ΔE) generated from classifying four classes (WM, GM, CSF and background) in MR brain volumes. Parameters for the Haar3D are 15^3 which is the maximum cube size that generates up to 44687 feature. (b) shows a histogram of Rectangle3D feature set scores (ΔE) generated from classifying two classes (femur and background) in 3D US images. Parameters for the rectangles are: maximum 3D rectangle dimensions is $(4,4,4)$ and maximum 3D search area around the 3D rectangle is $(x, y$ and z between $[-4,4]$). Number of features are $4^3 \times 9^3 = 46656$ while for instance the number of features which has $\Delta E \geq -0.6$ is 5171.

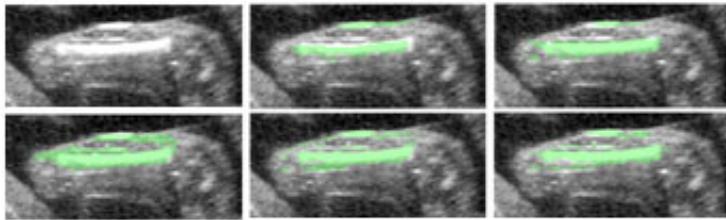


Fig. 2. Segmentation accuracy by different trees. Top left is a 2D fetal femur slice. The remaining five subfigures are the individual segmentation results of 5 separate trees.

3 Proposed Method

3.1 Training RF

Description of Feature sets. Several feature sets are constructed for a given image. We use the phrase "feature set" to denote the group of features of the same type but with different dimensions and locations around a Voxel Of Interest (VoxOI). We use Unary3D, Binary3D, Rectangle3D, Haar3D [10, 11], Position3D [2] and Averaged Rectangle3D which is the mean of the Rectangle3D feature set.

Offline Feature Selection. Since the feature pool contains many relatively "bad" features, see Fig. 1, the probability to select a good feature is low. This affects the creation of all trees where many nodes will poorly classify the data. If an RF classifier consists of T trees and each tree has an average height H then the number of candidate features to be tested is approximately $(2^H - 1) \times n' \times T$. This is a very time consuming process during training especially if the value of n' is large enough to guarantee the existence of a good feature within the n' ones. This is the main reason for using a small value of n' in all previous work [1-6]. We propose to exhaustively train all features on a well known gold standard (e.g., carefully manually segmented volume). The information gain (ΔE) is recorded as a performance for each feature. Subset of features from each feature set whose ΔE is greater than a specific value ($\Delta E_{\text{Threshold}}(\text{feature Set})$) are retained. The choice of such a threshold is feature set and application dependent, therefore has been chosen empirically in our work. From the score of each feature (ΔE), a smaller pool of features of size m (lookup table of size m) is chosen from the original pool (where $m \ll n$). The scores are normalized to a specific range since each feature set has a slightly different range. The training examples for this step can be a large enough subset of the whole training examples which has a well-defined ground truth. It is important that this subset captures all variations in the data. Although this process is time-consuming, it is required to be done only once for each dataset. For example, for a specific MRI scanner the offline feature selection step on brain images with their manual segmentation is performed once and the lookup table can then be used to efficiently train as many classifiers as possible. This means that this step is scanner specific and object-of-interest specific.

Training. The step can be done as in the classic RF by only considering features from the small feature pool of size m . However, we propose a different method which was inspired from [12] in which the trees were built totally in random (random feature and threshold with no training). Here, since we pre-trained all features and created a small feature pool of "good" features, we propose to use the pre-calculated score (ΔE) for each feature to create the trees. To create a decision node, m' out of m features are randomly selected. We choose the best feature out of m' depending on the pre-computed ΔE from the feature pool (lookup table). The chosen feature will then be applied to the training examples to construct a classifier (finding the best threshold). This means that we need m' lookup table comparisons and only one feature to compute at each tree node instead of n' in the standard RF technique. This makes the total number of features to compute approximately $[(2^H-1) \times 1 \times T]$ instead of $[(2^H-1) \times n' \times T]$ in the traditional RF framework i.e., saving a factor of n' where n' is typically >100 . The remaining of the training stage follows as in the classic RF technique. The main goals are to 1) enhance the choice of the chosen random feature and eliminate irrelevant ones and 2) significantly speed up the training process.

3.2 Testing RF

It is clear from Fig. 2 that each tree in a forest contributes differently to the final decision because of the random selection and creation of nodes. This also shows that classifiers vary in their classification accuracy. Therefore, we propose a weighted voting mechanism to calculate a probabilistic decision for each unseen example based on the strength of each tree to classify a specific unseen example. Remember that each node not only contains a classifier but also a score (ΔE) of how well that classifier did during the training stage. An unseen example follows a path on each starting from the root until it reaches a leaf. It approximately visits H (average tree height) classifiers on each tree until it reaches a leaf node. We used the mean of ΔE of the H classifiers as a measure of how good the tree is at classifying the test example. This can be formulated as,

$$\text{Score}(v, \text{tree}_t) = \frac{1}{H_t} \sum_{h=1}^{H_t} \Delta E_h (\text{node}_h \text{ at } \text{tree}_t). \quad (2)$$

After an unseen example is classified on T trees, the T scores are normalized so they sum to one,

$$\text{Norm}(\text{Score}(v, \text{tree}_t)) = \frac{1}{T-1} \left(1 - \frac{\frac{1}{H_t} \sum_{h=1}^{H_t} \Delta E_h (\text{node}_h \text{ at } \text{tree}_t)}{\sum_{i=1}^T \left(\frac{1}{H_i} \sum_{h=1}^{H_i} \Delta E_h (\text{node}_h \text{ at } \text{tree}_i) \right)} \right). \quad (3)$$

The normalized weight to classify an unseen example v is then embedded in the probabilistic classification so that the final decision becomes:

$$\forall c_j \in \{c_1, \dots, c_{\text{Labels}}\} \quad p(c_j | v, \text{RF}) = \sum_{t=1}^T \text{Norm}(\text{Score}(v, \text{tree}_t)) p(c_j | v, \text{leaf}(\text{tree}_t)). \quad (4)$$

4 Experimental Setup

Datasets. Two datasets were used to validate the proposed technique. One experiment used the IBSR dataset (<http://www.cma.mgh.harvard.edu/ibsr/>). The dataset consists of 20 T1-weighted MR brain images of normal subjects. The MR images were manually segmented to four objects namely White Matter (WM), Gray Matter (GM), Cerebro-Spinal Fluid (CSF) and background. The second experiment used 51 human fetal femur US volumes [13]. In this dataset, US volumes were acquired on 19 weeks fetuses ± 6 days. Volume dimensions are approximately $70 \times 70 \times 140$ voxels with a $(0.5 \times 0.5 \times 0.5)$ mm³ voxel spacing. Although segmenting the fetal femur in 3D US is not as popular as segmenting MRI brain, it is still of great clinical use and it is challenging. Several clinical studies [13, 14] have tried to correlate fetal bone growth to maternal characteristics especially maternal vitamin D. However, since US images have critical quality issues, they are good to prove the importance of the proposed feature selection and weighted voting to provide a more accurate 3D segmentation.

Validation Methodology. Experiments on traditional RF, pre-trained RF with weighted vote testing (Fast-Weighted RF (FWRF)) are reported. The validation was performed on the two datasets by comparing the manual and the automatic segmentations. In the first experiment, MR images were first bias field corrected using the N4 algorithm [15]. Since there were 20 cases, 5% of the training voxels were randomly selected from the 20 volumes for feature scoring and selection. Training was performed on 10 MR volumes while testing results were reported on the remaining 10 volumes. In the second experiment, from the 51 manually segmented cases, feature scoring and selection was performed on 10 3D fetal femoral US images. Training was done on another 20 volumes while results are reported on segmenting the femur from the remaining 21. If the number of training examples for each class in an RF framework are highly different, this can lead to highly imbalanced trees which in turn affects the classification accuracy. This is the case in both datasets. Therefore we randomly sub-sampled from the majority classes to approximately match the size of the minority class [16].

Implementation Details. In our experiments, the main training parameters are 1) the number of trees (T), 2) maximum tree depth (L), and 3) the number of features to test during node creation (n' in the classic RF and m' in FWRF). In the offline feature selection, the main parameters are ΔE threshold for each feature set which we set to keep approximately 10-15% of the original pool of the feature sets. All parameters were set experimentally and the main ones are ($T=20$, $L=15$, $n'=m'=200$) in the MRI experiment and ($T=20$, $L=13$, $n'=m'=100$) for fetal US. Since training and testing each tree is independent, parallel implementation is sensible. Therefore, we trained and tested trees in parallel. The parallelism usually drops the running time by at least a factor of $\min(T, \# \text{ of processors})$. In our experiments we used 10 processors due to limited computational power. Finally, the code was developed in the C# language.

5 Results

Brain Structures Segmentation in MR Images. Visual comparison between the classic RF and FWRF to segment brain MRI is shown in Fig. 3. On the other hand, Table 1 shows the Jaccard indices from several segmentation algorithms. Note that our

implementation of the classic RF versus the one implemented in [3] give slightly lower Jaccard index although the same dataset was used. The reason for this is that we could not reproduce [3] exactly as key parameters were not defined. On the other hand, the last row in the table shows better Jaccard indices over the classic RF. Finally, Table 2 shows the feature selection, training and testing times. Testing time is reported for segmenting one volume.

The Fetal Femur Segmentation in 3D US. We performed several comparisons in this experiment to validate the proposed technique and compare them with the conventional technique. Visual comparison is demonstrated in Fig. 4. The mean and standard deviation of precision, recall, dice coefficient and Jaccard index from all cases are reported in Table 4. Finally, feature selection, training and testing times are reported in Table 3.

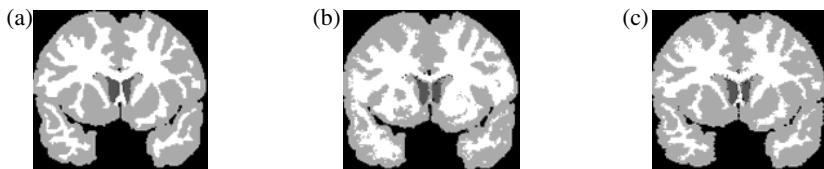


Fig. 3. Segmentation results on a brain MRI. (a) shows a 2D manually segmented slice from the 3D volume. (b) shows the segmentation using the classic RF. (c) shows the segmentation using FWRF. Black, dark gray, gray & white represent background, CSF, GM, and WM, respectively.

Table 1. Comparison between the proposed and several state of the art techniques. This table is based on [3]. Mean Jaccard indices are shown for segmenting CSF, GM and WM.

Method	CSF	GM	WM
Adaptive MAP	0.069	0.564	0.567
Biased MAP	0.071	0.558	0.562
Fuzzy c-means	0.048	0.473	0.567
Maximum-a-posteriori	0.071	0.550	0.554
Maximum-likelihood	0.062	0.535	0.551
Tree-Structure k-means	0.049	0.477	0.571
MPM-MAP	0.227	0.662	0.683
BSE/BFC/PVC	—	0.595	0.664
Constrained GMM	—	0.680	0.660
Spatial-varying GMM	—	0.768	0.734
Coupled surface	—	0.701	—
FSL	—	0.756	—
SPM	—	0.790	—
MAP with hist. [3]	0.549±0.02	0.814±0.00	0.710±0.01
Classic RF [3]	0.614±0.02	0.838±0.01	0.731±0.01
Classic RF	0.604±0.08	0.811±0.03	0.714±0.04
Fast-weighted RF	0.648±0.08	0.851±0.03	0.784±0.05

Table 2. Times to compute feature scores, training and testing for the classic RF and FWRF on the brain MRI dataset

	RF	FWRF
Feature Selection (hour)	0	26
Training (hour)	78	6
Testing (second)	11	23

Table 3. Times to compute feature scores, training and testing for the classic RF and FWRF on the 3D fetal femur US dataset

	RF	FWRF
Feature selection (h)	0	11
Training	16 h	95 min
Testing (second)	1.3	2.3

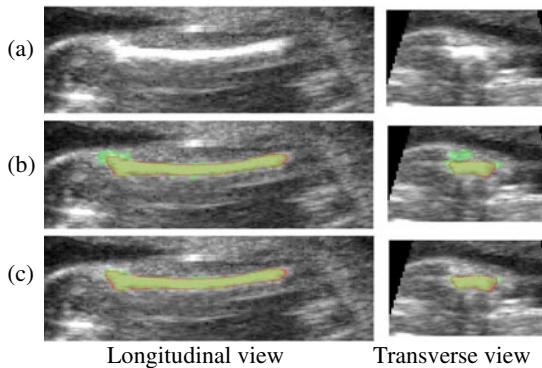


Fig. 4. Visual segmentation results on an US fetal femur. (a) shows a 2D US slice from the 3D volume. (b & c) show the segmentations from classic RF and FWRF respectively. In (b & c), the red boundary is an approximation of the manual segmentation; green voxels represent false negative which show voxels that were segmented by technique but not by the expert; the yellowish voxels represent the true positive which are the correctly segmented voxels.

Table 4. The mean \pm standard deviation of precision, recall, dice similarity and Jaccard index over 21 3D fetal femur US cases. comparisons are shown for classic RF, and FWRF.

	Precision	Recall	Dice	Jaccard
RF	0.68 \pm 0.18	0.89 \pm 0.11	0.75 \pm 0.11	0.61 \pm 0.13
FWRF	0.83 \pm 0.09	0.82 \pm 0.08	0.81 \pm 0.04	0.69 \pm 0.06

6 Discussion

Table 1 shows that FWRF outperforms the classic RF as well as many state of the art techniques. Two main contributions enhanced the classification accuracy over the classic RF. The feature selection step helped choose a smaller feature pool which contains mainly "good" features while weighted voting of trees produced a more accurate probabilistic decision.

FWRF has proven its applicability to segment different objects of interest in medical images. Moreover, we validated its performance on two image modalities. Although classic RF can be implemented in an efficient and parallel way, training is still time consuming. Therefore, we not only showed better classification accuracy in the FWRF but also much faster training. This is critical in some cases where training has to be done several times while tuning the main parameters. The feature selection is time-consuming and an extra overhead compared to the classic RF. Nevertheless, finding a global score for each feature is totally independent which results of a near-real time feature selection if computational power exists.

Extension to this work could be by enriching the feature pool with more features that can characterize different medical data accurately, e.g., HOG [17], local phase [18], etc. In addition, more interest is drawn recently in hybrid discriminative and generative models, e.g., [19], which makes our proposed technique a good candidate to combine with a generative model.

References

1. Criminisi, A., Shotton, J., Robertson, D., Konukoglu, E.: Regression forests for efficient anatomy detection and localization in CT studies. In: Menze, B., Langs, G., Tu, Z., Criminisi, A., et al. (eds.) MICCAI 2010. LNCS, vol. 6533, pp. 106–117. Springer, Heidelberg (2011)
2. Lempitsky, V., Verhoeek, M., Noble, J.A., Blake, A.: Random forest classification for automatic delineation of myocardium in real-time 3D echocardiography. In: Ayache, N., Delingette, H., Sermesant, M., et al. (eds.) FIMH 2009. LNCS, vol. 5528, pp. 447–456. Springer, Heidelberg (2009)
3. Yi, Z., Criminisi, A., Shotton, J., Blake, A.: Discriminative, semantic segmentation of brain tissue in MR images. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C., et al. (eds.) MICCAI 2009. LNCS, vol. 5762, pp. 558–565. Springer, Heidelberg (2009)
4. Geremia, E., Menze, B.H., Clatz, O., Konukoglu, E., Criminisi, A., Ayache, N.: Spatial Decision Forests for MS Lesion Segmentation in Multi-Channel MR Images. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6361, pp. 111–118. Springer, Heidelberg (2010)
5. Yaqub, M., et al.: Weighted Voting in 3D Random Forest Segmentation. In: MIUA (2010)
6. Criminisi, A., et al.: Decision Forests with Long-Range Spatial Context for Organ Localization in CT Volumes. In: MICCAI-PMMIA (2009)
7. Breiman, L.: Bagging predictors. *Mach. Learn.* 24(2), 123–140 (1996)
8. Breiman, L.: Random Forests. *Machine Learning* 45(1), 5–32 (2001)
9. Lepetit, V., Fua, P.: Keypoint recognition using randomized trees. *PAMI* 28(9), 1465–1479 (2006)
10. Shotton, J., et al.: Semantic Texton Forests for Image Categorization and Segmentation. In: CVPR (2008)
11. Oren, M., et al.: Pedestrian detection using wavelet templates. In: CVPR, pp. 193–199 (1997)
12. Geurts, P., et al.: Extremely randomized trees. *Mach. Learn.* 63(1), 3–42 (2006)
13. Mahon, P.A., et al.: The use of 3D ultrasound to investigate fetal bone development. *Norsk Epidemiologi*. 19(1) (2009)
14. Gale, C., et al.: Maternal vitamin D status during pregnancy and child outcomes. *EJCN* 62, 68–77 (2008)
15. Tustison, N.J., et al.: N4ITK: Improved N3 Bias Correction. *IEEE TMI* 29(6), 1310–1320 (2010)
16. Thomas, J., Jouve, P.-E., Nicoloayannis, N.: Optimisation and evaluation of random forests for imbalanced datasets. In: Esposito, F., Raś, Z.W., Malerba, D., Semeraro, G., et al. (eds.) ISMIS 2006. LNCS (LNAI), vol. 4203, pp. 622–631. Springer, Heidelberg (2006)
17. Dalal, N., Triggs, B.: Histograms of Oriented Gradients for Human Detection. In: CVPR, pp. 886–893 (2005)
18. Rajpoot, K., et al.: Local-phase based 3D boundary detection using monogenic signal and its application to real-time 3-D echocardiography images. In: ISBI (2009)
19. Tu, Z., et al.: Brain Anatomical Structure Segmentation by Hybrid Discriminative/Generative Models. *IEEE TMI*, 495–508 (2008)

Network-Based Classification Using Cortical Thickness of AD Patients

Dai Dai¹, Huiguang He¹, Joshua Vogelstein², and Zengguang Hou¹

¹ State Key Laboratory for Intelligent Control and Management of Complex Systems,
Institute of Automation, Chinese Academy of Sciences, Beijing, China

² Department of Applied Mathematics and Statistics, Johns Hopkins University, MD, USA
huiguang.he@ia.ac.cn

Abstract. In this article we propose a framework for establishing individual structural networks. An individual network is established for each subject using the mean cortical thickness of cortical regions as defined by the AAL atlas. Specifically, for each subject, we compute a similarity matrix of mean cortical thickness between pairs of cortical regions, which we refer to hereafter as the individual's network. Such individual networks can be used for classification. We use a combination of two types of feature selection approaches to search for the most discriminative edges. These edges serve as the input to a support vector machine (SVM) for classification. We demonstrate the utility of the proposed method by a comparison with classifying the raw cortical thickness data, and individual networks, using a publically available dataset. In particular, 83 subjects from the OASIS database were chosen to validate this approach, 39 of which were diagnosed with either mild cognitive impairment (MCI) or moderate Alzheimer's disease (AD) and the remaining were age-matched controls. While using an SVM on the raw cortical thickness data or individual networks without hybrid feature selection resulted in less than or nearly 80% classification accuracy, our approach yielded 90.4% classification accuracy in leave-one-out analysis.

1 Introduction

In recent years, studies related to brain connectivity are of increasing interest. Many researchers attempt to model functional connectivity through functional MRI, or establish anatomical networks through diffusion or structural MRI; these networks are analyzed using graph theory approaches to study their topological properties, which may help to interpret the operational mechanisms of human brain. In addition, some kinds of mental disorders (such as schizophrenia or Alzheimer's disease) have been discovered to be related with the abnormality of such topological properties [1,8]. These studies provide new clues to understand the pathology of the mental diseases. Moreover, they suggest that brain connectivity and its topological properties may be taken as biomarkers for computer-aided diagnosis.

Structural network construction can be grouped in two categories. In one category, a network is established for each group through, for instance, correlation analysis, and researchers focus mainly on the differences of two group-wise networks [1]. The authors of [1] calculate the Pearson correlation coefficient across subjects between

pairs of average gray matter volumes to obtain the connectivity matrix for each group. In the other category, individual network is built for each subject through, for example, fiber tracking. Such individual networks can be used for classification of individuals [3,5], or can be summarized to group networks for further analysis as well. [3] obtains a connectivity matrix for each subject through the robust L1-norm of mean cortical thickness and curvature. Graph edit distance and multidimensional scaling are used for feature extraction, and classification is performed via linear discriminant analysis. They obtain 0.93 area under ROC curve (AUC). In [4], the authors establish six networks for each MCI patient using six different parameters of DTI. An SVM-based method is used for feature selection and an SVM classifier is trained using the combination of these networks. They finally obtain 88.9% classification accuracy and AUC of 0.929 in the leave-one-out analysis.

In this paper, we present a method for establishing individual structural networks and attempt to use these networks to distinguish AD patients from normal controls. To obtain a connectivity matrix for each subject, we first calculate the distance of pairs of cortical regions using mean cortical thickness; then we apply a kernel function to the distance to obtain the connection weight. For feature selection, a combination of two kinds of method is used: (i) a filter method for fast and rough dimensionality reduction and (ii) a wrapper method for further precise feature selection. Classification is performed via SVM. We apply our method to 83 subjects selected from OASIS AD database to evaluate the performance.

This paper makes several contributions. First, we demonstrate the utility of a network-based classification using merely cortical thickness data. Second, we present a method for establishing individual networks. Third, we develop a fast and precise feature selection approach that improves classification performance. The results of this paper suggest several possible avenues for future research, including alternate kernels for establishing individual networks and new dimensionality reduction techniques.

2 Materials and Methods

2.1 Data Acquisition

Data used in this study are taken from Open Access Series of Imaging Studies (OASIS) database (www.oasis-brains.org/). The OASIS database consists of a cross-sectional collection of 416 right-handed subjects aged from 18 to 96, in which one hundred subjects over 60 years old have been clinically diagnosed with very mild to moderate AD. In our study, 83 subjects are chosen from the database for experiment, including 39 subjects with MCI and moderate AD (22 females, 17 males, age \pm SD = 77.77 \pm 5.64), and 44 age-matched normal controls (30 females, 14 males, age \pm SD = 75.77 \pm 7.29). The control subjects have a CDR (Clinical Dementia Rating) of zero, and MMSE (Mini-Mental State Examination) scores between 26 and 29. The MCI/AD patients have a CDR of 0.5, 1 or 2 and MMSE scores between 15 and 28 (most of patients have MMSE scores less than 26). We divide the subjects into two groups based on their CDR scale, that is, all subjects with CDR \geq 0.5 are in class 0, while all subjects with CDR = 0 are in class 1.

Cortical thickness is measured using the method described in [5,6]. Briefly, the registered images are first classified into gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) using an artificial neural network classifier [5]. Then the inner (GM/WM) and outer (pial) cortical surfaces are extracted automatically through the CLASP algorithm [6]. Each surface consists of 40,962 vertices and 81,920 mesh triangles in each hemisphere. Cortical thickness is calculated as the Euclidean distance between the linked vertices on the two surfaces.

2.2 Construction of Individual Networks

A network, or graph, is typically defined as $G = (V, E)$, where V is the set of vertices (or nodes) and E is the set of edges (or links). For this work, we assume a node is a cortical region as defined by the AAL atlas. Thus each individual's network, or graph, shares the same set of 80 labeled vertices (non-surface regions of AAL atlas are excluded). This facilitates performing comparisons using only the edges. We assume each edge is the “similarity” between a pair of nodes, as measured by the kernel described below. Therefore, the similarity matrix is symmetric and with ones along the diagonal. More specifically, mean cortical thickness of each cortical region (vertex) is calculated, obtaining a $(n \times V)$ matrix, where n is the number of subjects and V is the number of vertices. Let $T_k(i)$ (or $T_k(j)$) denote the cortical thickness of the i -th (or j -th) ROI of the k -th subject; then the connection weight $w_k(i, j)$ is defined as follow,

$$w_k(i, j) = k(d_k(i, j)) = \exp\left(-\frac{d_k(i, j)}{\sigma}\right) \quad (1)$$

where $k(\cdot)$ is a kernel function, and $d_k(i, j)$ is a distance function. Specifically, the exponential kernel $k(d) = \exp(-d/\sigma)$ is used here, which contains an input parameter σ that determines how locally data is analyzed, and the distance function $d_k(i, j)$ is defined as,

$$d_k(i, j) = [T_k(i) - T_k(j)]^2 \quad (2)$$

Other kernel function and distance function could also be used to construct the networks. We chose exponential kernel because this weight function is non-negative and monotonically decreasing, and is easy to interpret: a larger distance results in smaller similarity and vice versa. We set the kernel width $\sigma = 0.015$ which made the edge weight of most individual networks nearly distributed uniformly between zero and one. According to the above method, we obtained n individual connectivity matrices with dimensionality of $(V \times V)$. These networks are pruned to reduce noise: edges with weight less than 0.01 are eliminated.

2.3 Classification of Individual Networks

We attempt to use these individual networks for classification. The input data are $n = 83$ connectivity matrices and each network has mostly $p = V \times (V - 1)/2 = 3160$ edges, yielding a $(n \times p)$ feature matrix for classification. Because of the high-dimensionality of this problem, dimensionality reduction is considered to reduce

the variance and improve the performance of classifier. Because we are interested in interpretable results, we consider only canonical dimensionality reduction, a.k.a., feature selection (as opposed to feature extraction, in which new dimensions are constructed as functions of the original dimensions). Feature selection algorithms can be grossly subdivided into two categories: filter methods and wrapper methods [7]. A filter method directly evaluates feature subsets through their information content while a wrapper method iteratively optimizes the performance of a specific classification algorithm. Thus filter methods tend to be more computationally efficient; they also tend to be less effective than wrapper methods. Hence, a hybrid approach is intuitively desirable. In our study, a two sample t-test acts to filter the features roughly by retaining all features with p-values smaller than a set threshold ($p\text{-value} < 0.05$), and a local-learning-based method [2] is implemented on the remaining feature subset for further feature selection (the parameters of this method is chosen under the author's recommend). The latter method outputs a score for each feature; we only retain features with scores larger than 0.1. Then a SVM with the radial basic function (RBF) kernel is trained using the selected feature subset. In order to avoid overfitting and obtain an unbiased results, we apply a nested leave-one-out cross validation (LOOCV) to evaluate the performance of our method (for more detail of nested cross validation, see [11]). Specifically, two nested loops of cross validation (CV) are conducted. In the inner loop, the best hyperparameters of SVM are selected according to the inner CV accuracy, using only the training set ($n - 1$ subjects); while in the outer loop, the selected SVM model is evaluated on the test set (the remaining one subject). We report the outer CV accuracy as the final classification accuracy.

3 Results

3.1 Classification Result of Individual Networks

The classification accuracy of LOOCV analysis is 90.4% (92.3% sensitivity and 88.6% specificity) using the feature subset selected by the hybrid feature selection method. The receiver operator characteristic (ROC) curve can be found in Fig.1 and the area under ROC curve (AUC) is 0.9487. This performance on OASIS database suggests that it is an effective method for AD classification.

To further demonstrate the effect of the proposed method (IN - Hybrid), we compared the results of (i) classifying raw cortical thickness data without feature selection (RCT - None), (ii) classifying raw cortical thickness data with local-learning-based feature selection (RCT - LLB), and (iii) classifying individual networks using only 2-sample t-test ($p\text{-value} < 0.05$) for feature selection (IN - tTest), via a SVM. The classification performance of different descriptions of data in LOOCV analysis is

Table 1. Classification performance and AUC for different descriptions of data

	<i>Accuracy</i>	<i>Sensitivity</i>	<i>Specificity</i>	<i>AUC</i>
RCT - None ⁽ⁱ⁾	74.7 %	71.8 %	77.3 %	0.8176
RCT - LLB ⁽ⁱⁱ⁾	80.7 %	74.4 %	86.4 %	0.8310
IN - tTest ⁽ⁱⁱⁱ⁾	68.7 %	60.0 %	77.3 %	0.7552
IN - Hybrid ^(Proposed)	90.4 %	92.3 %	88.6 %	0.9487

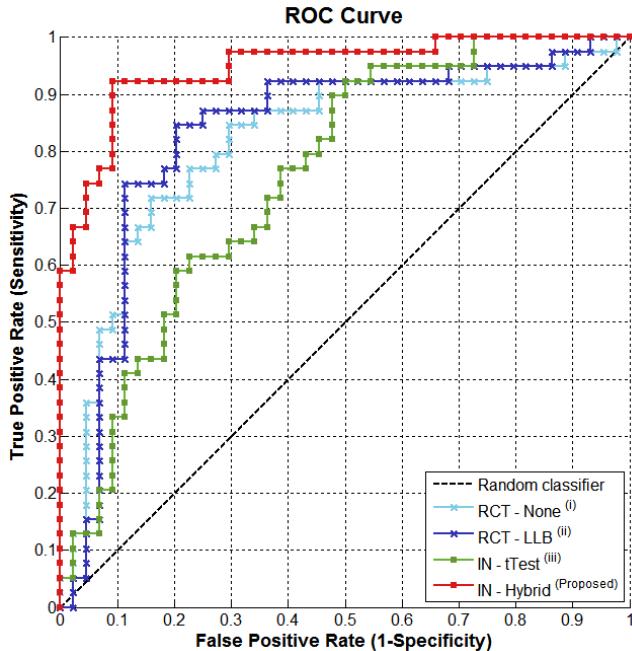


Fig. 1. ROC curve for leave-one-out analysis

summarized in Table 1, and Fig. 1 shows the comparison of their ROC curves. We find that the classification directly using cortical thickness outperforms that using individual networks with only 2 sample t-test feature selection; this might be caused by the high dimensionality of the network features which can reduce the performance of most classification algorithm including SVM. Therefore a hybrid feature selection in conjunction with classification was applied to this problem, improving classification accuracy by approximately 22%. Moreover, the performance of our method is also better than that using raw cortical thickness data, suggesting that networks might be a better representation of the intrinsic structure of cortical data.

3.2 Most Discriminative Features Selected by the Proposal Method

The number of features (or edges) selected by the proposed method in leave-one-out analysis ranges from 10 to 15; in total, 22 different edges were selected in at least one of the 83 folds of cross validation. We calculate the reproducibility ratio of each feature selected across LOOCV iterations to find which of them are almost consistent or chosen in most feature selections out of the 83 folds of cross validation, because these consistent features might encode the class-conditional signal. In Table 2 we list some of edges that are frequently selected in LOOCV analysis and point out the cortical regions they connect. We find the selected edges connect several regions that are believed to be related with Alzheimer's disease [1,4,9,10,12], including hippocampus¹, anterior and middle cingulum², superior temporal gyrus³, fusiform gyrus⁴, precuneus⁵, lingual gyrus⁶, calcarine⁷, Supramarginal gyrus⁸, and several regions in the frontal lobes such as orbitofrontal region⁹; these regions are marked with the numbers above in Table 2.

We also apply the proposed feature selection method to all of the 83 individual networks and have 11 edges chosen, all of which are included in the 22 edges in LOOCV analysis and marked with an asterisk in Table 2. Note that these 11 edges are just those occur most frequently in LOOCV analysis (except the 327th feature which has a score of 0.21, slightly larger than 0.1, while others all have a score larger than 0.8); this may suggest that the proposed feature selection method is robust. In Fig.2 we show the 11 most discriminative edges with red (or blue) lines connected to the related regions in the average brain of the 83 subjects.

Table 2. Features selected in LOOCV analysis that have most discriminative ability. Only those that are selected more than 10 times are listed in the table. Features with asterisk on their right side are the 11 features that are selected when using all the subjects for analysis. The abbreviations of AAL regions can be found in [1].

<i>Index of feature</i>	<i>Ratio of Reproducibility</i>	<i>Represented edge</i>	
607*	100 %	ORBmid.L ⁹	IFGoper.R
697*	100 %	ORBmid.R ⁹	ACG.R ²
868*	100 %	IFGoper.R	PCUN.R ⁵
1414*	100 %	OLF.L	LING.L ⁶
2460*	100 %	CUN.L	LING.R ⁶
2795*	100 %	FFG.L ⁴	PCUN.R ⁵
3047*	100 %	PCUN.L ⁵	STG.R ³
632*	98.8 %	ORBmid.L ⁹	HIP.L ¹
996*	98.8 %	IFGtriang.R	SMG.L ⁸
2446*	98.8 %	CAL.R ⁷	HES.L
1986	49.4 %	ACG.R ²	DCG.R ²
327*	41.0 %	ORBsup.L ⁹	OLF.R
930	41.0 %	IFGtriang.L	SMG.L ⁸
2346	25.3 %	PHG.R ¹	LING.R ⁶
747	18.1 %	IFGoper.L	IFGtriang.L
681	12.0 %	ORBmid.R ⁹	ORBinf.R ⁹

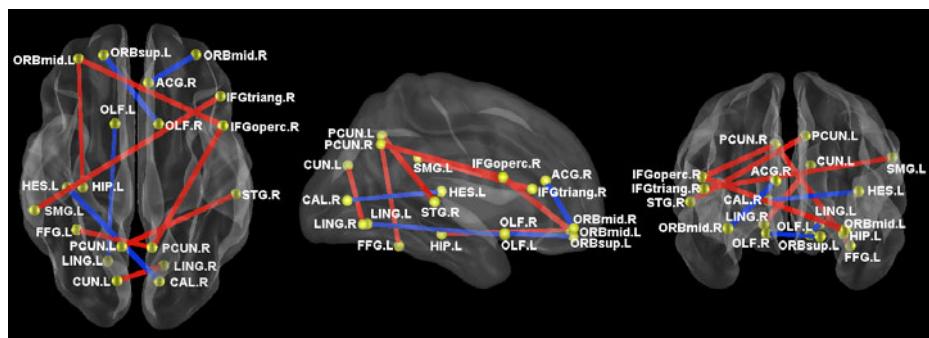


Fig. 2. A visualization of the 11 edges that have most discriminative power, which are selected by the proposed feature selection method when using all of the 83 subjects. The red (or blue) line indicates that the average weight of edges in normal group is larger (or smaller) than that in patient group.

4 Conclusion

We have proposed an approach for establishing individual cortical networks, as well as a joint feature selection and classification method for classifying individual networks. For network building, we calculate the distance between pairs of cortical regions using mean cortical thickness via a simple distance function, and then obtain the connection weight of each network through an exponential kernel function. A different definition of distance and kernel function might result in networks with distinct characteristics, which may reveal the structure of raw cortical data from different perspectives. Combining filter and wrapper feature selection method results in a precise and computationally efficient algorithm; and the classification performance of our method on OASIS database suggests that the proposed method is very effective for AD classification. Moreover, the selected features are interpretable and are consistent with previous neurobiological findings in AD patients using other methods. We find networks could represent the intrinsic structure of cortical data better, and network-based classification might be a promising approach for computer-aided diagnosis.

In future work, we plan to examine and compare the classification performance of individual networks established by different distance and kernel functions; and more effective algorithms for feature selection and network-based classification might be introduced or developed. In addition, a topological analysis of the individual networks via graph theory approach will be conducted to check for any abnormalities of topological properties.

Acknowledgement. This work was supported by the National Natural Science Foundation of China (20670530, 60875079), Beijing Nova Plan (2007A094), and the Sci. & Tech. Aiding the Disabled Program of the Chinese Academy of Sciences (Grant #KGCX2-YW-618).

References

1. Yao, Z., Zhang, Y., Lin, L., Zhou, Y., Xu, C., Jiang, T.: Abnormal Cortical Network in Mild Cognitive Impairment and Alzheimer's disease. *PLoS Computational Biology* 6(11), e1001006 (2010)
2. Sun, Y., Todorovic, S., Goodison, S.: Local-Learning-Based Feature Selection for High-Dimensional Data Analysis. *IEEE Trans. PAMI* 32(9), 1610–1626 (2010)
3. Raj, A., Mueller, S.G., Young, K., Laxer, K.D., Weiner, M.: Network-level analysis of cortical thickness of the epileptic brain. *NeuroImage* 52(4), 1302–1313 (2010)
4. Wee, C.Y., Yap, P.T., Denny, K., Browndyke, J.N., Potter, G.G., Welsh-Bohmer, K.A., Wang, L., Shen, D.: Enriched white-matter connectivity networks for accurate identification of MCI patients. *NeuroImage* 54(3), 1812–1822 (2011)
5. Zijdenbos, A., Forghani, R., Evans, A.: Automatic quantification of MS lesions in 3D MRI brain data sets: Validation of INSECT. In: Wells, W.M., Colchester, A.C.F., Delp, S.L. (eds.) MICCAI 1998. LNCS, vol. 1496, pp. 439–448. Springer, Heidelberg (1998)
6. Kim, J.S., Singh, V., Lee, J.K., Lerch, J., Ad-Dab'bagh, Y., MacDonald, D., Lee, J.M., Kim, S.I., Evans, A.C.: Automated 3-D extraction and evaluation of the inner and outer cortical surfaces using a Laplacian map and partial volume effect classification. *NeuroImage* 27(1), 210–221 (2005)

7. Kohavi, R., John, G.H.: Wrappers for feature subset selection. *Artificial Intelligence* 97(1/2), 273–324 (1997)
8. Liu, Y., Liang, M., Zhou, Y., He, Y., Hao, Y., Song, M., Yu, C., Liu, H., Liu, Z., Jiang, T.: Disrupted small-world networks in schizophrenia. *Brain* 131(4), 945–961 (2008)
9. Chetelat, G., Landeau, B., Eustache, F., Mezenge, F., Viader, F., de La Sayette, V., Desgranges, B., Baron, J.C.: Using voxel-based morphometry to map the structural changes associated with rapid conversion in MCI: a longitudinal MRI study. *NeuroImage* 27(4), 934–946 (2005)
10. Karas, G.B., Scheltens, P., Rombouts, S., Visser, P.J., Van Schijndel, R.A., Fox, N.C., Barkhof, F.: Global and local gray matter loss in mild cognitive impairment and Alzheimer's disease. *Neuroimage* 23(2), 708–716 (2004)
11. Wilson, S.M., Ogar, J.M., Laluz, V., Growdon, M., Jang, J., Glenn, S., Miller, B.L., Weiner, M.W., Gorno-Tempini, M.L.: Automated MRI-based classification of primary progressive aphasia variants. *Neuroimage* 47(4), 1558–1567 (2009)
12. Bozzali, M., Parker, G.J.M., Serra, L., Embleton, K., Gili, T., Perri, R., Caltagirone, C., Cercignani, M.: Anatomical connectivity mapping: A new tool to assess brain disconnection in Alzheimer's disease. *Neuroimage* 54(3), 2045–2051 (2010)

Anatomical Regularization on Statistical Manifolds for the Classification of Patients with Alzheimer's Disease

Rémi Cuingnet^{1,2}, Joan Alexis Glaunès^{1,3}, Marie Chupin¹,

Habib Benali², and Olivier Colliot¹

The Alzheimer's Disease Neuroimaging Initiative*

¹ Université Pierre et Marie Curie-Paris 6, CNRS UMR 7225, Inserm UMR_S 975,
Centre de Recherche de l'Institut Cerveau-Moelle (CRICM), Paris, France

² Inserm, UMR_S 678, LIF, Paris, France

³ MAP5, Université Paris 5 - René Descartes, Paris, France

Abstract. This paper introduces a continuous framework to spatially regularize support vector machines (SVM) for brain image analysis based on the Fisher metric. We show that, by considering the images as elements of a statistical manifold, one can define a metric that integrates various types of information. Based on this metric, replacing the standard SVM regularization with a Laplace-Beltrami regularization operator allows integrating to the classifier various types of constraints based on spatial and anatomical information. The proposed framework is applied to the classification of magnetic resonance (MR) images based on gray matter concentration maps from 137 patients with Alzheimer's disease and 162 elderly controls. The results demonstrate that the proposed classifier generates less-noisy and consequently more interpretable feature maps with no loss of classification performance.

1 Introduction

Brain image analyses have widely relied on univariate voxel-wise analyses, such as voxel-based morphometry (VBM) for structural MRI [1]. In such analyses, brain images are first spatially registered to a common stereotaxic space, and then mass univariate statistical tests are performed in each voxel to detect significant group differences. However, the sensitivity of these approaches is limited when the differences are spatially complex and involve a combination of different voxels or brain structures [2].

Recently, there has been a growing interest in support vector machines (SVM) methods [3, 4] to overcome the limits of these univariate analyses. These approaches allow capturing complex multivariate relationships in the data and have

* Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (adni.loni.ucla.edu). A complete listing of ADNI investigators can be found at: http://adni.loni.ucla.edu/wp-content/uploads/how_to_apply/ADNI_Authorship_List.pdf

been successfully applied to the individual classification of a variety of neurological and psychiatric conditions such as Alzheimer's disease [5, 6, 7, 8, 9] fronto-temporal dementia [5], schizophrenia [10] and Parkinsonian syndromes [11]. Moreover, the output of the SVM can also be analyzed to localize spatial patterns of discrimination, for example by drawing the coefficients of the optimal margin hyperplane (OMH) – which, in the case of a linear SVM, live in the same space as the MRI data [6, 7].

However, voxel-based comparisons are subject to registration errors and inter-individual variability. Therefore, one of the problems with analyzing directly the OMH coefficients is that the corresponding maps are scattered and lack spatial coherence. This makes it difficult to give a meaningful interpretation of the maps, for example to localize the brain regions altered by a given pathology. This is due to the fact that the regularization term of the standard linear SVM is not a spatial regularization. To overcome this limitation, Cuingnet et al. [12] proposed to directly enforce spatial consistency into the SVM by using the Laplacian of a regularization graph. They proposed a regularization graph which takes into consideration both spatial information (the location) and anatomical information (the tissue types). They combine spatial and anatomical information by modifying the local topology induced by the spatial information with respect to some given anatomical priors (tissue types). Since the images are discrete, they used a discrete framework to model local behaviors: graphs. Nevertheless, as the brain is intrinsically a continuous object, it seems more interesting to describe local behaviors from the continuous viewpoint.

This paper extends this spatial regularization framework to the continuous case. In particular, we show that by considering images as statistical manifolds together with the Fisher metric, it allows taking into account various prior information such as tissue, atlas information and spatial proximity. We then apply the proposed framework to the classification of MR images based on gray matter concentration maps and cortical thickness measures from patients with Alzheimer's disease and elderly controls. The results demonstrate that the proposed approach allows obtaining spatially and anatomically coherent discrimination patterns. It generates more interpretable features maps with an increase or at least with no loss of classification performance.

2 Spatially Regularized SVM on Riemannian Manifold

2.1 Background

In this contribution, we consider the case of brain images which are spatially normalized to a common stereotaxic space as in many group studies or classification methods [6, 7, 9, 10, 13]. These images can be any characteristics extracted from the MRI, such as tissue concentration maps (in VBM). Let $(\mathbf{x}_s)_{s \in [1, N]}$ be the images of N subjects and $(y_s)_{s \in [1, N]} \in \{\pm 1\}^N$ their group labels (e.g. diagnosis). For each subject s , \mathbf{x}_s can be considered as a square integrable real-valued function defined on a compact subset, \mathcal{V} , of \mathbb{R}^3 or more generally on a compact

of a 3D Riemannian manifold. Let \mathcal{V} be the domain of the 3D images. SVMs search for the hyperplane for which the margin between groups is maximal. The standard linear SVM solves the following optimization problem [3, 4]:

$$(\mathbf{w}^{\text{opt}}, b^{\text{opt}}) = \arg \min_{\mathbf{w} \in L^2(\mathcal{V}), b \in \mathbb{R}} \frac{1}{N} \sum_{s=1}^N \ell_{\text{hinge}}(y_s [\langle \mathbf{w}, \mathbf{x}_s \rangle_{L^2} + b]) + \lambda \|\mathbf{w}\|_{L^2}^2 \quad (1)$$

where $\lambda \in \mathbb{R}^+$ is the *regularization parameter* and ℓ_{hinge} the *hinge loss function* defined as: $\ell_{\text{hinge}} : u \in \mathbb{R} \mapsto (1 - u)^+$.

With a linear SVM, the *feature space* is the same as the *input space*. Thus, when the input features are images, the weight map \mathbf{w}^{opt} is also an image. This map qualitatively informs us about the role of the different brain regions in the classifier [9]. Therefore, since two neighboring regions should have a similar role in the classifier, \mathbf{w}^{opt} should be smooth with respect to the topology of \mathcal{V} . However, this is not guaranteed with the standard linear SVM because the regularization term *is not a spatial regularization*.

2.2 Regularization Operator

By considering the SVM from the regularization viewpoint [4], one can constrain \mathbf{w}^{opt} to be smooth with respect to the topology of \mathcal{V} . This is done through the definition of a *regularization operator*, P , defined as a linear map from a space $\mathcal{U} \subset L^2(\mathcal{V})$ into $L^2(\mathcal{V})$. When P is bijective and symmetric,

$$\min_{\mathbf{u} \in \mathcal{U}, b \in \mathbb{R}} \frac{1}{N} \sum_{s=1}^N \ell_{\text{hinge}}(y_s [\langle \mathbf{u}, \mathbf{x}_s \rangle_{L^2} + b]) + \lambda \|P\mathbf{u}\|_{L^2}^2 \quad (2)$$

is equivalent to a linear SVM on the data $(P^{-1}\mathbf{x}_s)_s$. Similarly, it can be seen as a SVM minimization problem on the raw data with kernel K defined by $K(\mathbf{x}_1, \mathbf{x}_2) = \langle P^{-1}\mathbf{x}_1, P^{-1}\mathbf{x}_2 \rangle_{L^2}$. One has to define the regularization operator P to obtain the suitable regularization for the problem.

2.3 Spatial Regularization on Compact Riemannian Manifold

Spatial regularization requires the notion of proximity between elements of \mathcal{V} . In this paper, \mathcal{V} is considered as a 3-dimensional compact Riemannian manifold (\mathcal{M}, g) with boundaries. The metric, g , then models the notion of proximity. On such spaces, the heat kernel exists [14, 15]. Therefore, the Laplacian regularization presented in [12] can be extended to compact Riemannian manifolds.

Let Δ_g denotes the Laplace-Beltrami operator¹. Let $(\mathbf{e}_n)_{n \in \mathbb{N}}$ be an orthonormal basis of $L^2(\mathcal{V})$ of eigenvectors of Δ_g (with homogeneous Dirichlet boundary conditions) [14, 16] and $(\mu_n)_{n \in \mathbb{N}}$ the corresponding eigenvalues. We define \mathcal{U}_β

$$\mathcal{U}_\beta = \left\{ \mathbf{u} = \sum_{n \in \mathbb{N}} u_n \mathbf{e}_n \mid (u_n)_{n \in \mathbb{N}} \in \ell^2 \text{ and } \left(e^{\frac{1}{2}\beta\mu_n} u_n \right)_{n \in \mathbb{N}} \in \ell^2 \right\}$$

¹ Note that, with the convention used in this paper, in Euclidean space, $\Delta_g = -\Delta$ where Δ is the Laplacian operator.

where ℓ^2 denotes the set of square-summable sequences. We chose the regularization operator $P_\beta : \mathcal{U}_\beta \rightarrow L^2(\mathcal{V})$ defined as:

$$P_\beta : \mathbf{u} = \sum_{n \in \mathbb{N}} u_n \mathbf{e}_n \mapsto e^{\frac{1}{2}\beta \Delta_g} \mathbf{u} = \sum_{n \in \mathbb{N}} e^{\frac{1}{2}\beta \mu_n} u_n \mathbf{e}_n \quad (3)$$

This penalizes the high-frequency components with respect to the topology of \mathcal{V} .

3 Spatial Proximity

When the proximity is encoded by a Euclidean distance, this is equivalent to pre-process the data with a Gaussian smoothing kernel with standard deviation $\sigma = \sqrt{\beta}$. However such a metric does not take into account anatomical information. In this section, the goal is to define a metric that takes into account various prior informations such as tissue, atlas and location information. We first show that this can be done by considering the images as elements of a statistical manifold and using the Fisher metric. We then give some details about the computation of the Gram matrix.

3.1 Fisher Metric

The images are registered to a common space. Therefore, when considering some location $v \in \mathbb{R}^3$, the true location is known up to the registration errors. Such spatial information can be modeled by a probability density function: $x \in \mathbb{R}^3 \mapsto p_{\text{loc}}(x|v)$. A simple example would be $p_{\text{loc}}(\cdot|v) \sim \mathcal{N}(v, \sigma_{\text{loc}}^2)$. It can be seen as a confidence index about the spatial localization at voxel v .

We further assume that we are given an anatomical or a functional atlas \mathcal{A} composed of R regions: $\{\mathcal{A}_r\}_{r=1 \dots R}$. Therefore, in each point $v \in \mathcal{V}$, we have a probability distribution $p_{\text{atlas}}(\cdot|v) \in \mathbb{R}^{\mathcal{A}}$ which informs us about the atlas region in v . As a result, in each point $v \in \mathbb{R}^3$, we have some information about the spatial location and some anatomical information through the atlas. Such information can be modeled by a probability density function $p(\cdot|v) \in \mathbb{R}^{\mathcal{A} \times \mathbb{R}^3}$. Therefore, we consider the parametric family of probability distributions:

$$\mathcal{M} = \left\{ p(\cdot|v) \in \mathbb{R}^{\mathcal{A} \times \mathbb{R}^3} \right\}_{v \in \mathcal{V}}$$

In the following, we further assume that p_{loc} and p_{atlas} are independent. Thus, p verifies: $p((\mathcal{A}_r, \mathbf{x})|v) = p_{\text{atlas}}(\mathcal{A}_r|v)p_{\text{loc}}(\mathbf{x}|v), \forall (\mathcal{A}_r, \mathbf{x}) \in \mathcal{A} \times \mathbb{R}^3$. We also assume that p is sufficiently smooth in $v \in \mathcal{V}$ and that the Fisher information matrix is definite at each $v \in \mathcal{V}$. Then the parametric family of probability distributions \mathcal{M} can be considered as a differential manifold [17]. A natural way to encode proximity on \mathcal{M} is to use the Fisher metric, since such metric is invariant under reparametrization of the manifold. \mathcal{M} with the Fisher metric is a compact Riemannian manifold [17]. The metric tensor g is then given for all $v \in \mathcal{V}$ by:

$$g_{ij}(v) = \mathbb{E}_v \left[\frac{\partial \log p(\cdot|v)}{\partial v_i} \frac{\partial \log p(\cdot|v)}{\partial v_j} \right], \quad 1 \leq i, j \leq 3$$

When $p_{\text{loc}}(\cdot|v) \sim \mathcal{N}(v, \sigma_{\text{loc}}^2 I_3)$, we have: $g_{ij}(v) = g_{ij}^{\text{atlas}}(v) + \frac{\delta_{ij}}{\sigma_{\text{loc}}^2}$.

3.2 Computing the Gram Matrix

The computation of the kernel matrix requires the computation of $e^{-\beta \Delta_g} \mathbf{x}_s$ for all the subjects of the training set. The eigendecomposition of the Laplace-Beltrami operator is intractable since the number of voxels in a brain images is about 10^6 . Hence $e^{-\beta \Delta_g} \mathbf{x}_s$ is considered as the solution at time $t = \beta$ of the heat equation with the Dirichlet homogeneous boundary conditions of unknown \mathbf{u} :

$$\frac{\partial \mathbf{u}}{\partial t} + \Delta_g \mathbf{u} = 0; \quad \mathbf{u}(t = 0) = \mathbf{x}_s \quad (4)$$

To solve equation (4), one can use a variational approach [18]. We used the rectangular finite elements $\{\phi^{(i)}\}$ in space and the explicit finite difference scheme for the time discretization. Δ_x and Δ_t denote the space step and the time step respectively. Let $U(t)$ denote the coordinates of $\mathbf{u}(t)$. Let U^n denote the coordinates of $\mathbf{u}(t = n\Delta_t)$. This leads to:

$$\mathbf{M} \frac{dU}{dt}(t) + \mathbf{K}U(t) = 0; \quad U(t = 0) = U^0 \quad (5)$$

$$\text{with } \mathbf{K}_{i,j} = \int_{\mathcal{V}} \left\langle \nabla_{\mathcal{M}} \phi^{(i)}, \nabla_{\mathcal{M}} \phi^{(j)} \right\rangle_{\mathcal{M}} d\mu_{\mathcal{M}} \text{ and } \mathbf{M}_{i,j} = \int_{\mathcal{V}} \phi^{(i)} \phi^{(j)} d\mu_{\mathcal{M}} \quad (6)$$

where \mathbf{K} is the stiffness matrix and \mathbf{M} is the mass matrix.

The explicit finite difference scheme was used for the time discretization, thus U^{n+1} is given by: $\mathbf{M}U^{n+1} = (\mathbf{M} - \Delta_t \mathbf{K})U^n$. The step Δ_x is fixed by the MRI spatial resolution. The time step, Δ_t , is then chosen so as to respect the Courant-Friedrichs-Lowy (CFL) condition: $\Delta_t \leq 2(\max \lambda_i)^{-1}$ where λ_i are the eigenvalues of the general eigenproblem: $\mathbf{K}\mathbf{U} = \lambda\mathbf{M}\mathbf{U}$. Therefore, the computational complexity is: $O(N\beta(\max_i \lambda_i)d)$. To compute the optimal time step Δ_t , we estimated the largest eigenvalue with the power iteration method. In our experiments, for $\sigma_{\text{loc}} = 5$, $\lambda_{\text{max}} \approx 15.4$ and for $\sigma_{\text{loc}} = 10$, $\lambda_{\text{max}} \approx 46.5$.

3.3 Setting the Diffusion Parameter β

Our method required the tuning of two parameters σ_{loc} and β . The parameter σ_{loc} was chosen a priori. As evaluating the spectrum of the Laplacian operator is intractable considering the images' sizes, β was chosen to be equivalent to the diffusion parameter of the Gaussian smoothing, $\beta = \sigma^2$, where σ is the standard deviation for the Gaussian smoothing kernel. To be comparable with the Euclidean case, we first normalized g with:

$$\left(\frac{1}{|\mathcal{V}|} \int_{u \in \mathcal{V}} \frac{1}{3} \text{tr} \left(g^{\frac{1}{2}}(u) \right) du \right)^2$$

4 Experiments and Results

In this section, the proposed framework is applied to the analysis of MR images using gray matter concentration maps from patients with Alzheimer's disease and elderly controls.

4.1 Materials

Subjects and MRI acquisition. Data used in the preparation of this article were obtained from the Alzheimer’s Disease Neuroimaging Initiative (ADNI) database. The Principal Investigator of this initiative is Michael W. Weiner, MD, VA Medical Center and University of California, San Francisco. ADNI is the result of efforts of many co-investigators from academic institutions and private corporations. For up-to-date information, see www.adni-info.org.

We used the same study population as in [9]. As a result, 299 subjects were selected: 162 cognitively normal elderly controls (76 males, 86 females, age \pm SD [range] = 76.3 ± 5.4 [60 – 90] years, and mini-mental score (MMS) = 29.2 ± 1.0 [25 – 30]) and 137 patients with AD (67 males, 70 females, age = 76.0 ± 7.3 [55 – 91] years, and MMS = 23.2 ± 2.0 [18 – 27]). The T1-weighted MR images described in [19] were used in this study.

Features Extraction. All images were segmented into gray matter (GM), white matter (WM) and cerebrospinal fluid (CSF) using the SPM5 unified segmentation routine [20] and spatially normalized using the DARTEL diffeomorphic registration algorithm [21] with the default parameters. The features are the modulated GM probability maps in the MNI space.

4.2 Classification Experiments

We tested the spatial regularization for both the Euclidean metric and the Fisher metric. In the following, they will be referred to as *Regul-Euclidean* and *Regul-Fisher* respectively. The atlas information used was only the tissue types (GM, WM and CSF templates). To assess the impact of the regularization we also performed the classification experiments with no regularization: *Direct*.

Optimal coefficient maps. The optimal SVM weights \mathbf{w}^{opt} for different value of β are shown on Figure 1. When no spatial regularization has been carried out (a), the \mathbf{w}^{opt} maps are noisy and scattered. With Euclidean spatial regularization (b-c), they become smoother and more spatially consistent. However it mixes tissues and does not respect the topology of the cortex. With the Fisher metric (d-e), the obtained map is much more consistent with the brain anatomy. Compared to the Euclidean regularization, it better respects the topology of the cortex (Fig. 2). The main regions in which atrophy increases the likelihood of being classified as AD (regions in red) are: the medial temporal lobe, the inferior and middle temporal gyri, the posterior cingulate and the posterior middle frontal gyri.

Classification performances. In order to obtain unbiased estimates of the performances, the set of participants was randomly split into two groups of the same size: a training set and a testing set. On the training set, a gridsearch with a leave-one-out-cross-validation was used to estimate the optimal values of the hyperparameters: the cost parameter C ($\lambda = \frac{1}{2NC}$) of the linear C-SVM

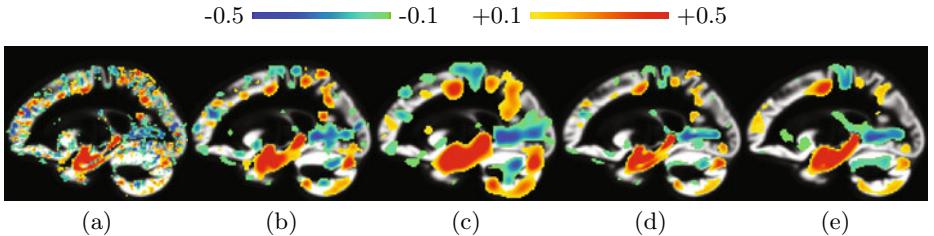


Fig. 1. Normalized w^{opt} coefficients for: (a) *Direct*, (b-c) *Regul-Euclidean* with FWHM = 4 mm and FWHM = 4 mm respectively, (d-e) *Regul-Fisher* with FWHM \sim 4 mm and FWHM \sim 8 mm respectively ($\sigma_{\text{loc}} = 10$). In all experiments, $C = 1$.

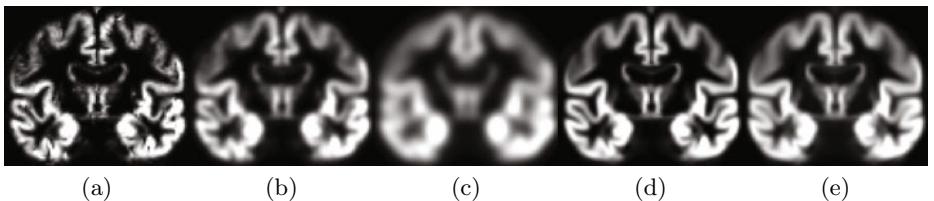


Fig. 2. Gray probability map ((a) original map) of a control subject preprocessed with: (b) a 4 mm FWHM gaussian kernel, (c) an 8 mm FWHM gaussian kernel, (d)-(e) with $e^{-\frac{\beta}{2}\Delta_g}$ and β corresponds to a 4 mm and to an 8 mm FWHM respectively.

$(10^{-5}, 10^{-4.5}, \dots, 10^3)$, FWHM $(0, 2, \dots, 8 \text{ mm})$ and σ_{loc} $(5, 10 \text{ mm})$. The performances of the resulting classifiers were then evaluated on the testing set. Classification performances in terms of accuracies were slightly improved by spatially regularizing the SVM with the Fisher metric: *Direct*: 89%, *Regul-Euclidean*: 89%, *Regul-Fisher* : 91%, *COMPARE* [10]: 86%, *STAND-Score* [7]: 81%.

5 Conclusion

In conclusion, this paper presents a continuous framework to spatially regularize SVM for brain image analysis based on the Fisher metric. By considering the images as elements of a statistical manifold, one can define a metric that integrates various types of information. Based on this metric, replacing the standard SVM regularization with a Laplace-Beltrami regularization operator allows integrating to the classifier various types of constraints based on spatial and anatomical information. The proposed approach makes the results more consistent with the anatomy, making their interpretation more meaningful. Finally, it should be noted that the proposed approach is not specific to structural MRI, and can be applied to other pathologies and other types of data (e.g. functional or diffusion-weighted MRI).

Acknowledgements. This work was supported by ANR (project HM-TC, number ANR-09-EMER-006).

Data collection and sharing for this project was funded by the Alzheimer's Disease Neuroimaging Initiative (ADNI) (National Institutes of Health Grant U01 AG024904).

References

1. Ashburner, J., Friston, K.J.: Voxel-based morphometry—the methods. *NeuroImage* 11(6), 805–821 (2000)
2. Davatzikos, C.: Why voxel-based morphometric analysis should be used with great caution when characterizing group differences. *NeuroImage* 23(1), 17–20 (2004)
3. Vapnik, V.N.: *The Nature of Statistical Learning Theory*. Springer, Heidelberg (1995)
4. Schölkopf, B., Smola, A.J.: *Learning with Kernels*. MIT Press, Cambridge (2001)
5. Davatzikos, C., et al.: Individual patient diagnosis of AD and FTD via high-dimensional pattern classification of MRI. *NeuroImage* 41(4), 1220–1227 (2008)
6. Klöppel, S., et al.: Automatic classification of MR scans in Alzheimer's disease. *Brain* 131(3), 681–689 (2008)
7. Vemuri, P., et al.: Alzheimer's disease diagnosis in individual subjects using structural mr images: Validation studies. *NeuroImage* 39(3), 1186–1197 (2008)
8. Gerardin, É., et al.: Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. *NeuroImage* 47(4), 1476–1486 (2009)
9. Cuingnet, R., et al.: Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage* 56(2), 766–781 (2011)
10. Fan, Y., et al.: COMPARE: classification of morphological patterns using adaptive regional elements. *IEEE Transactions on Medical Imaging* 26(1), 93–105 (2007)
11. Duchesne, S., et al.: Automated computer differential classification in Parkinsonian syndromes via pattern analysis on MRI. *Academic Radiology* 16(1), 61–70 (2009)
12. Cuingnet, R., Rosso, C., Lehéricy, S., Dormont, D., Benali, H., Samson, Y., Colliot, O.: Spatially regularized SVM for the detection of brain areas associated with stroke outcome. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) *MICCAI 2010. LNCS*, vol. 6361, pp. 316–323. Springer, Heidelberg (2010)
13. Querbes, O., et al.: Early diagnosis of Alzheimer's disease using cortical thickness: impact of cognitive reserve. *Brain* 132(8), 2036–2047 (2009)
14. Jost, J.: *Riemannian geometry and geometric analysis*. Springer, Heidelberg (2008)
15. Lafferty, J., Lebanon, G.: Diffusion kernels on statistical manifolds. *JMLR* 6, 129–163 (2005)
16. Hebey, E.: *Sobolev spaces on Riemannian manifolds*. Springer, Heidelberg (1996)
17. Amari, S.I., et al.: *Differential Geometry in Statistical Inference*, vol. 10. Institute of Mathematical Statistics (1987)
18. Druet, O., Hebey, E., Robert, F.: Blow-up theory for elliptic PDEs in Riemannian geometry. Princeton Univ. Press, Princeton (2004)
19. Jack, C.R., et al.: The Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging* 27(4) (2008)
20. Ashburner, J., Friston, K.J.: Unified segmentation. *NeuroImage* 26(3), 839–851 (2005)
21. Ashburner, J.: A fast diffeomorphic image registration algorithm. *NeuroImage* 38(1), 95–113 (2007)

Rapidly Adaptive Cell Detection Using Transfer Learning with a Global Parameter

Nhat H. Nguyen¹, Eric Norris², Mark G. Clemens², and Min C. Shin¹

¹ Department of Computer Science, University of North Carolina, Charlotte

² Department of Biology, University of North Carolina, Charlotte

{fnhnguye1, enorris9, mgclemen, mcshing}@unc.edu

Abstract. Recent advances in biomedical imaging have enabled the analysis of many different cell types. Learning-based cell detectors tend to be specific to a particular imaging protocol and cell type. For a new dataset, a tedious re-training process is required. In this paper, we present a novel method of training a cell detector on new datasets with minimal effort. First, we combine the classification rules extracted from existing data with the training samples of new data using transfer learning. Second, a global parameter is incorporated to refine the ranking of the classification rules. We demonstrate that our method achieves the same performance as previous approaches with only 10% of the training effort.

1 Introduction

Dramatic advances in biological imaging over the past decade, including both the development of more powerful imaging hardware and novel fluorescent probes, have revolutionized many areas of biological research [9]. The manual approach of cell analysis is too tedious and error-prone; thus it is not feasible for handling large datasets. Automated cell detection is a growing field of interest with a wide range of applications which permits statistical analysis of various cell parameters such as apoptosis, adherence, morphology and motility [1]. Thus, it has the potential to identify even subtle effects of many physiological stimuli on many cell types [4].

Recently, a number of automated cell detection methods, which are based on machine learning algorithms, have been proposed [2, 4, 8]. To our knowledge, these cell detection methods are not effective on different cell types since they required a large number of training samples from each cell type. For each new cell dataset with different appearance (e.g., size, shape, color), the users often need to re-train the algorithm by collecting training samples. This is a tedious and time-consuming process. Thus, there is a great need for a method that can be rapidly trained on new datasets with minimal training effort.

In this paper, we propose a novel cell detection method that can be rapidly trained to new datasets. The goal is to minimize the number of samples required to train the detection method which should translate to the reduction of human effort. We use a transfer learning algorithm [10] to leverage the classification rules gathered from existing data (source classes) to improve detection on the new data (target class). By incorporating the cell size distribution as a global

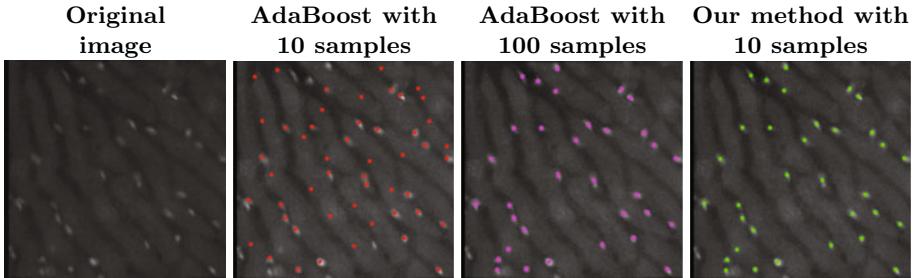


Fig. 1. Sample detection results from AdaBoost and our method. Note the poor performance of AdaBoost with a small training effort (10 samples). The proposed method with only 10 training samples is able to achieve equal performance of AdaBoost with 100 training samples.

parameter on new dataset, we further increase the accuracy of the detection algorithm using only a minimal number of training samples. The cell size distribution is determined during the training step and does not need to be re-trained for each individual image. We refer to our method as the GlobalTrAdaBoost, a boosting-based method that integrates a global parameter into the transfer learning framework. The evaluation on five cell types with 50 real images (2660 cells) demonstrates that the GlobalTrAdaBoost is able to achieve equal performance of a typical pattern recognition algorithm with only 10% of the training effort required. Sample detection results are shown in Figure 1.

2 Method

We first describe a baseline cell detection using Adaptive Boosting. Then, we explain the proposed method of using transfer learning, followed by the inclusion of the global parameter. The overview of the method is shown in Figure 2.

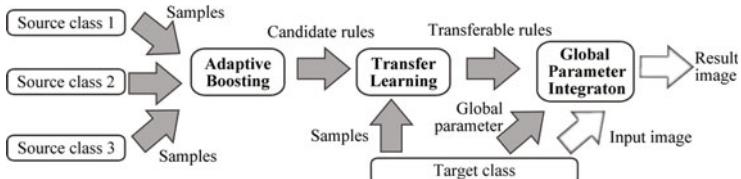


Fig. 2. The overview of the proposed method. The gray arrow indicates the training step while the white arrow indicates the testing step.

2.1 Adaptive Boosting (AdaBoost)

Boosting is an iterative method of constructing an accurate classifier by combining many weak classifiers, each of which only needs to be reasonably accurate [3].

One of the most popular boosting methods, Adaptive Boosting (also known as AdaBoost), can be used to train the cell detector.

The AdaBoost algorithm weights each weak classifier based on its prediction accuracy. In the feature space \mathcal{X} , and the label space $\mathcal{Y} = \{-1, +1\}$ denoting a pixel as background or cell sample, the detection task is to estimate a classifier function $f : \mathcal{X} \rightarrow \mathcal{Y}$ given training data $D = \{(\mathbf{x}_i, y_i) | \mathbf{x}_i \in \mathcal{X}, y_i \in \mathcal{Y}, 1 \leq i \leq n\}$ where \mathbf{x}_i and y_i are respectively the feature vector and the label of training sample i , and n is the number of training samples. Initially, AdaBoost constructs a distribution of weights $\mathbf{w} = \{w_i | w_i = \frac{1}{n}, 1 \leq i \leq n\}$ over the training data. For each iteration $t = 1$ to T , AdaBoost selects a weak classifier that gives the least classification error as $h_t(\mathbf{x}_i) = \arg \min(\epsilon_t)$ where $\epsilon_t = \sum_i w_i [y_i \neq h_t(\mathbf{x}_i)]$. In the next iteration, the weights associated with the samples misclassified by the selected weak classifier are increased as

$$w_i \leftarrow w_i e^{-\alpha_t y_i h_t(\mathbf{x}_i)} \quad (1)$$

where $\alpha_t = \frac{1}{2} \ln \frac{1-\epsilon_t}{\epsilon_t}$. Finally, the strong classifier \hat{f} is computed as the signum function of the weighted linear combination of T weak classifiers

$$\hat{f} = \text{sign}\left(\sum_t \alpha_t h_t(\mathbf{x}_i)\right). \quad (2)$$

2.2 Transfer Learning with AdaBoost (TaskTrAdaBoost)

Since AdaBoost requires a large number of training samples to be effective, training with only few samples often leads to poor performance. Transfer learning re-uses the classification rules from source classes with the target class to minimize the required amount of target training samples. We use a recent transfer learning algorithm, TaskTrAdaBoost, to conduct the training process [10].

First, we would like to gather the classification rules (weak classifiers) from the source classes. Let source class S_k ($1 \leq k \leq K$ where K is the number of source classes) contain the source training data $D^{S_k} = \{(\mathbf{x}_j, y_j) | 1 \leq j \leq n^{S_k}\}$ where n^{S_k} is the number of source training samples. We use AdaBoost from Section 2.1 to train the classifier function \hat{f}^{S_k} based on D^{S_k} . Then, we collect all candidate weak classifier h_c from \hat{f}^{S_k} of K source classes into set $\mathcal{H}_c = \{h_c(\mathbf{x}_j) | h_c(\mathbf{x}_j) \in \hat{f}^{S_k}, 1 \leq c \leq C\}$ where $C = T \times K$.

Second, we use set \mathcal{H}_c to build the strong classifier \hat{f}^τ for the target class \mathcal{T} given training data $D^\tau = \{(\mathbf{x}_l, y_l) | 1 \leq l \leq n^\tau\}$ where n^τ is the number of target training samples (note that $n^\tau \ll n^{S_k}$). The target weight distribution is initialized as $\mathbf{w}^\tau = \{w_l | w_l = \frac{1}{n^\tau}, 1 \leq l \leq n^\tau\}$. For each iteration, we find a transferable weak classifier h_β that minimizes the error over target data D^τ as

$$h_\beta(\mathbf{x}_l) = \arg \min_{h_\beta \in \mathcal{H}_c} (\epsilon_l) \quad (3)$$

where the classification error is computed as: $\epsilon_l = \sum_l w_l [y_l \neq h_\beta(\mathbf{x}_l)]$.

With few training samples n^τ , it is possible that more than one h_β yield the same minimal error ϵ_l over D^τ . Since the TaskTrAdaBoost uses only one

h_β , it could select the h_β which over-fits the training samples and thus reduces performance on the target class. In the next section, we propose to use a global parameter, distribution of cell sizes, to refine the ranking of weak classifiers to regularize the problem and alleviate the risk of over-fitting.

2.3 Global Parameter Integration (GlobalTrAdaBoost)

We select the transferable weak classifier h_β to conform with a global parameter, the cell size distribution, in addition to minimize the target training error. We measure the conformity of h_β to the cell size distribution using the Kullback-Leibler divergence, a non-symmetric measure of the difference between a true distribution and an approximated model [5]. Under *in vivo* microscopy, the fluorescent intensity within the cell population differs depending on the location of a cell with respect to the microscope focal lens resulting in different cell sizes [6]. Thus, we model the cell sizes using a Gaussian distribution.

Let us define the cell size distribution $\mathcal{P} \sim \mathcal{N}(\mu_m, \sigma_m^2)$ with $\mu_m = \frac{1}{M} \sum_m (r_m)$, $\sigma_m = \sqrt{\frac{1}{M} \sum_m (r_m - \mu_m)^2}$ where r_m is the cell size and $1 \leq m \leq M$ is the number of size samples. In other words, the cell size distribution \mathcal{P} can be estimated using set $\mathcal{R} = \{r_m | 1 \leq m \leq M\}$. In the target training image \mathcal{I}_τ , a user can measure r_m by one additional mouse click on the cell boundary after getting the cell location. Thus, acquiring M samples of cell sizes requires only additional M mouse clicks. In Section 3.3, we show that the estimation of \mathcal{P} is robust enough to maintain stable performance with $M = 6$.

After acquiring \mathcal{P} , we collect all h_β into set $\mathcal{H}_\beta = \{h_\beta(\mathbf{x}_l) | 1 \leq \beta \leq B\}$ where B is the number of weak classifiers that satisfy (3). Then, we employ each $h_\beta \in \mathcal{H}_\beta$ to classify the target training image \mathcal{I}_τ to obtain a binary classification image containing cell and background pixels. Using the connected component labeling procedure, we group cell pixels into cell regions and construct set $\mathcal{R}_\beta = \{r_u | 1 \leq u \leq U\}$ where r_u is the radius of a cell region and U is the number of detected regions. From set \mathcal{R}_β , we can compute the detected cell size distribution $\mathcal{Q}_\beta \sim \mathcal{N}(\mu_u, \sigma_u^2)$. We select the transferable weak classifier h'_β which minimizes the Kullback-Leibler divergence between \mathcal{P} and \mathcal{Q}_β :

$$h'_\beta(\mathbf{x}_l) = \arg \min_{h_\beta \in \mathcal{H}_\beta} (\mathcal{D}_{KL}(\mathcal{P} || \mathcal{Q}_\beta)) \quad (4)$$

where $\mathcal{D}_{KL}(\mathcal{P} || \mathcal{Q}_\beta) = \sum_p \mathcal{P} \log \frac{\mathcal{P}(p)}{\mathcal{Q}_\beta(p)}$ and p is the bin containing the range of cell size values [5]. As the result, the weak classifier h'_β conforms with the global parameter \mathcal{P} besides minimizing the target training error. The procedures to update \mathbf{w}^τ for each iteration and to construct the strong classifier \hat{f}^τ are similar to (1) and (2), respectively.

3 Experiments

3.1 Evaluation Procedure

Data Description. Five different cell types (white blood cells, natural killer T-cells, HT29 colon cancer, red blood cells, and drosophila) are acquired using

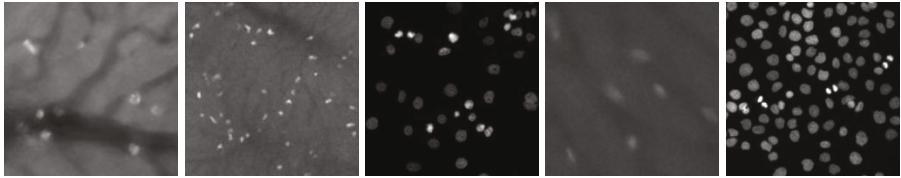


Fig. 3. Representative images of five different cell types (from left to right): white blood cells, natural killer T-cells, HT29 colon cancer, red blood cells, and drosophila

2 imaging protocols (*in vivo* epi-fluorescence and isolated fluorescently labeled) and 3 magnification levels (10X, 20X, and 40X). Sample images of each cell types are shown in Figure 3. We evaluate the performance of the detection algorithms on a total of 50 real images (10 images from each cell type). Each image contains from 25 to 100 cells (total of 2660 cells). We divide the images from each cell type into two halves for training and evaluating. A biology technician has manually determined the center and the radius of the cells in 50 images.

Local Features. The feature space \mathcal{X} is defined by a 10-D local feature vector \mathbf{x}_i from the training sample i . The description of each feature is explained as follow. First, we compute the normalized radial mean response of sample i as the ratio between the mean intensity of the inner circle to the outer circle surrounding the pixel location of i [7]. We apply six different scales for the inner and outer circle diameters to accommodate a variety of cell sizes. Second, we calculate the mean of gradient magnitude within a square region around sample i . The edge length of the region is same as the largest circle diameter from above. Third, we collect the filtering responses at sample i from circular averaging, low-pass Gaussian and isotropic Laplacian of Gaussian kernels. The response values are normalized by dividing with the maximum within the entire image. Note that additional features can be added for potential improvement.

Performance Metric. We measure the performance according to the manually marked ground truth data. Each cell detection by an algorithm is determined as true positive if there is a corresponding ground truth cell within the mean value of cell radii r_m (estimated by the user, as discussed in Section 2.3). We compute the number of true positives (TP), false positives (FP) and false negatives (FN). Recall and precision are computed as $\frac{TP}{TP+FN}$ and $\frac{TP}{TP+FP}$, respectively. For measuring the overall performance, we use F-measure, a harmonic mean of precision and recall, as $F = 2 \times \frac{Precision \times Recall}{Precision + Recall}$.

3.2 Detection Accuracy

To evaluate the detection accuracy, each cell type is chosen as a target class, and the remaining cell types are used as source classes. We measure the training effort E_τ as the number of target training samples and the number of size samples required to obtain the global parameter ($E_\tau = n_\tau + M$). Note that $M = 6$ for

Table 1. F-Measures (Mean \pm Standard Error of the Mean) of boosting-based cell detection methods for different training efforts. A performance number is highlighted in bold if it is significantly better than other methods based on a paired t-test at $p = 0.05$. $Global_{under}$ and $Global_{over}$ (as discussed in Section 3.3) are two versions of GlobalTrAdaBoost with fluctuated values of the cell size distribution.

E_τ	AdaBoost	TaskTrAdaBoost	GlobalTrAdaBoost	$Global_{under}$	$Global_{over}$
10	0.57 ± 0.024	0.69 ± 0.018	0.81 ± 0.014	0.78 ± 0.032	0.79 ± 0.038
20	0.68 ± 0.019	0.72 ± 0.018	0.82 ± 0.011	0.79 ± 0.039	0.80 ± 0.042
30	0.69 ± 0.018	0.75 ± 0.015	0.81 ± 0.011	0.79 ± 0.031	0.80 ± 0.012
40	0.76 ± 0.013	0.78 ± 0.012	0.81 ± 0.008	0.81 ± 0.034	0.81 ± 0.008
50	0.76 ± 0.012	0.77 ± 0.012	0.82 ± 0.006	0.82 ± 0.023	0.82 ± 0.023
60	0.79 ± 0.010	0.80 ± 0.010	0.83 ± 0.008	0.82 ± 0.008	0.81 ± 0.043
70	0.80 ± 0.009	0.82 ± 0.007	0.83 ± 0.004	0.82 ± 0.007	0.82 ± 0.014
80	0.80 ± 0.012	0.81 ± 0.010	0.83 ± 0.006	0.83 ± 0.008	0.82 ± 0.011
90	0.80 ± 0.012	0.81 ± 0.009	0.83 ± 0.004	0.83 ± 0.009	0.83 ± 0.010
100	0.82 ± 0.009	0.83 ± 0.007	0.84 ± 0.004	0.83 ± 0.008	0.83 ± 0.010

GlobalTrAdaBoost and $M = 0$ for AdaBoost and TaskTrAdaBoost. Training is conducted with E_τ varying from 10 to 100. For each value of E_τ , we conduct 30 executions of training and testing on each of 5 target classes. In each execution, the target training samples D_τ are randomly selected. The performances (in terms of F-measures) are shown in Table 1.

When trained with a large number of samples ($E_\tau = 100$), AdaBoost, TaskTrAdaBoost, and GlobalTrAdaBoost reach similar maximum performance (F-measures equal to 0.82, 0.83, and 0.84, respectively). However, GlobalTrAdaBoost shows significant improvement to other methods with small training efforts. First, at training effort $E_\tau = 10, 20$, and 30 , the average improvement of the proposed method over TaskTrAdaBoost is 17%, 14%, and 8% with standard error reduced by 22%, 39%, and 27%, respectively (see Table 1). Second, GlobalTrAdaBoost only needs 10% of the training effort ($E_\tau = 10$ versus 100) to achieve the same performance ($p = 0.87$) as AdaBoost. Third, with just 10 training samples, GlobalTrAdaBoost's F-measure is already 0.81, which is only 4% lower than the maximum performance. The sample results comparing AdaBoost and the proposed method are shown in Figure 4.

3.3 Sensitivity of Global Parameter

In this section, we investigate the sensitivity of the algorithm's performance with respect to the accuracy of the global parameter estimation. To fully estimate the range of values that a global parameter can take for a dataset, we randomly select sets of M size samples. For each set of M samples, we compute the mean and standard deviation of cell sizes. We repeat this randomly sampling process 30 times in each dataset. We observe that the mean of the standard deviations of multiple sets start converging when $M \geq 6$. For each cell type, we examine multiple sets of 6 size samples and compute the mean μ_Δ and standard deviation σ_Δ . If the GlobalTrAdaBoost performance is still higher than other methods

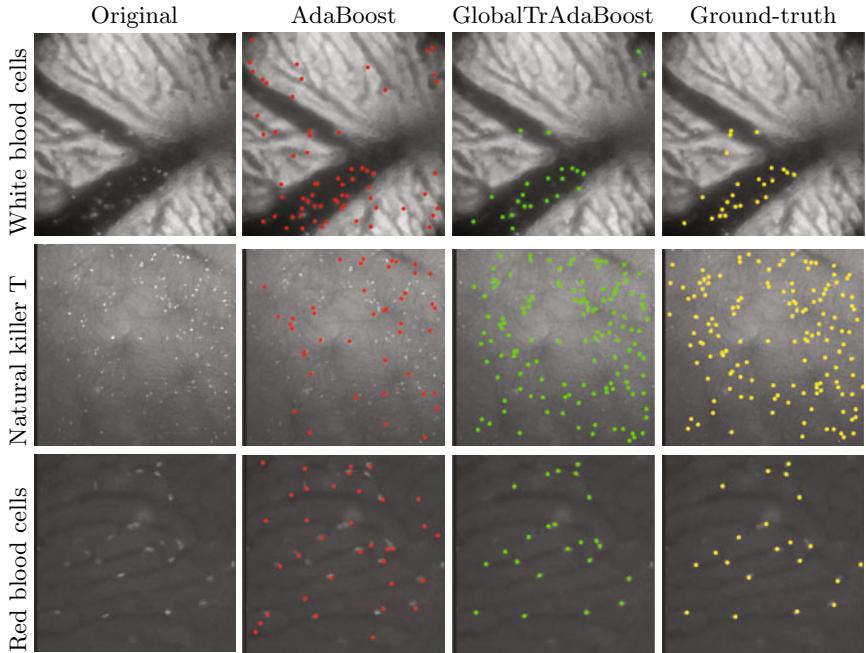


Fig. 4. Sample results comparing AdaBoost and the proposed method (GlobalTrAdaBoost) using 10 training samples

when the global parameter estimation varies by the value of σ_Δ , then 6 cell size samples ($M = 6$) are sufficient to estimate the cell size distribution.

Consequently, we integrate the GlobalTrAdaBoost with two fluctuated values of the global parameter $\mathcal{P}_{under} \sim \mathcal{N}(\mu_\Delta - \sigma_\Delta, \sigma_\Delta^2)$, and $\mathcal{P}_{over} \sim \mathcal{N}(\mu_\Delta + \sigma_\Delta, \sigma_\Delta^2)$. The corresponding methods $Global_{under}$ and $Global_{over}$ are executed 30 times in the same procedure described in Section 3.2. We show the F-measures in conjunction with other detection methods in Table 1. The performance of both $Global_{under}$ and $Global_{over}$ at each E_τ are significantly better ($p < 0.05$) than both AdaBoost and TaskTrAdaBoost up to $E_\tau = 60$.

4 Conclusion

In this paper, we integrate a global parameter into the transfer learning algorithm to reduce the amount of training effort required for cell detection. Our method is able to achieve the performance of previous boosting-based algorithms with only 10% of the training effort. To further improve our algorithm, we plan to incorporate additional global features to handle overlapping cells and investigate in the automated estimation of the global parameter. We believe that these results demonstrate the potential of the proposed method for greater applicability in cell detection by reducing the amount of manual effort.

References

- [1] Balagopalan, L., Sherman, E., Barr, V., Samelson, L.: Imaging techniques for assaying lymphocyte activation in action. *Nat. Rev. Immunol.* 11(1), 21–33 (2011)
- [2] Carpenter, A., Jones, T., Lamprecht, M., Clarke, C., Kang, I., Friman, O., Guertin, D., Moffat, J., et al.: Cellprofiler: image analysis software for identifying and quantifying cell phenotypes. *Genome Biology* 7(10), R100 (2006)
- [3] Freund, Y., Schapire, R.: A short introduction to boosting. *Japonese Society for Artificial Intelligence* 14(5), 771–780 (1999)
- [4] Hodneland, E., Bukoreshtliev, N., Eichler, T., Tai, X.C., Gurke, S., Lundervold, A., Gerdes, H.H.: A unified framework for automated 3-d segmentation of surface-stained living cells and a comprehensive segmentation evaluation. *IEEE Transactions on Medical Imaging* 28(5), 720–738 (2009)
- [5] Kullback, S.: The kullback-leibler distance. *The American Statistician* 41(4), 340–341 (1987)
- [6] Mukherjee, D., Ray, N., Acton, S.: Level set analysis for leukocyte detection and tracking. *IEEE Transactions on Image Processing* 13(4), 562–572 (2004)
- [7] Nguyen, N., Keller, S., Huynh, T., Shin, M.: Tracking colliding cells. In: *IEEE Workshop in Applications of Computer Vision* (2009)
- [8] Pan, J., Kanade, T., Chen, M.: Heterogeneous conditional random field: Realizing joint detection and segmentation of cell regions in microscopic images. In: *2010 on Computer Vision and Pattern Recognition (CVPR)*, pp. 2940–2947 (2010)
- [9] Toomre, D., Bewersdorf, J.: A new wave of cellular imaging. *Annual Review of Cell and Developmental Biology* (January 2010)
- [10] Yao, Y., Doretto, G.: Boosting for transfer learning with multiple sources. In: *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 1855–1862 (June 2010)

Automatic Morphological Classification of Lung Cancer Subtypes with Boosting Algorithms for Optimizing Therapy

Ching-Wei Wang¹ and Cheng-Ping Yu²

¹Graduate Institute of Biomedical Engineering, National Taiwan University of Science and Technology, Taiwan

²Department of Pathology, Tri-Service General Hospital, Taiwan
cweiwang@mail.ntust.edu.tw

Abstract. Patient-targeted therapies have recently been highlighted as important. An important development in the treatment of metastatic non-small cell lung cancer (NSCLC) has been the tailoring of therapy on the basis of histology. A pathology diagnosis of “non-specified NSCLC” is no longer routinely acceptable; an effective approach for classification of adenocarcinoma (AC) and squamous carcinoma (SC) histotypes is needed for optimizing therapy. In this study, we present a robust and objective automatic classification system for real time classification of AC and SC based on morphological tissue pattern of H&E images alone to assist medical experts in diagnosis of lung cancer. Various original and extended Densitometric and Haralick’s texture features are used to extract image features, and a Boosting algorithm is utilized to train the classifier, together with alternative decision tree as the base learner. For evaluation, 369 tissue samples were collected in tissue microarray format, including 97 adenocarcinoma and 272 squamous carcinoma samples. Using 10-fold cross validation, the technique achieved high accuracy of 92.41%, and we also found that the two Boosting algorithms (cw-Boost and AdaBoost.M1) perform consistently well in comparison with other popularly adopted machine learning methods, including support vector machine, neural network, single decision tree and alternative decision tree. This approach offers a robust, objective and rapid procedure for optimized patient-targeted therapies.

Keywords: morphological classification, computer vision, adenocarcinoma, squamous carcinoma, boosting, tissue microarray.

Introduction

Lung cancer (LC) is the leading cause of cancer-related death worldwide and accounts for over one million deaths annually [4]. The majority of patients with NSCLC present with locally advanced or metastatic disease for which systemic treatment with chemotherapy is the standard of care. Those patients with earlier stage disease (I and II predominantly) are eligible for surgery with curative

intent. NSCLC is a highly drug resistant cancer and response rates are poor, particularly in the second line setting where response rate is less than 10%. LC is classified on the basis of histology into two main subtypes. These are non-small cell lung cancer (NSCLC), which represents around 80-90%, and small cell lung cancer (SCLC). For NSCLC, five year survival is around 5-10% and has not changed for over two decades. NSCLC is sub-divided into squamous carcinoma (SC) and non-squamous histological subtypes, and the non-squamous includes adenocarcinoma (AC) and undifferentiated carcinoma. AC accounts for about 40% of all lung cancers, whereas SC accounts for 25% to 30% of all lung cancers [15]. Overall, AC and SC account for 65% to 70% of all lung cancer patients.

An important development in the treatment of metastatic NSCLC has been the tailoring of therapy for patients with NSCLC on the basis of histology. In the recent JMRB phase III clinical trial [4], pemetrexed with cisplatin has been shown to improve survival in patients with AC but has adverse effect on SC patients. SC tumours are contraindicated for treatment with antiangiogenesis agents where tumour related bleeding can lead to mortality, and FDA has restricted the use of pemetrexed to non-SC NSCLC patients only. In contrast, SC patients fared better with the gemcitabine combination. Separation of efficacy based on histology has also been shown for bevacizumab in the ECOG 1499 study [13] in which patients with AC only, benefited from the combination with paclitaxel-carboplatin. Conversely, early data on IGF inhibitors suggests that SC are more susceptible compared with AC. Accordingly, optimizing therapy for NSCLC in the first line setting now warrants robust, rapid and economically efficient approaches for classification of adenocarcinoma and squamous cell carcinoma types for therapeutic purposes.

A pathology diagnosis of "non-specified NSCLC" is no longer routinely acceptable, and a robust and efficient approach for classification of AC and SC histotypes is needed for optimizing therapy. However, due to complex tissue patterns, to the authors' best knowledge, there is no automated method in classification of AC and SC based on morphological tissue patterns of H&E images. Ullmann et al. [16] presented an approach to classify AC and SC using a number of biomarker expression based on immunohistochemistry; the tissue expression levels of 86 different proteins were manually scored by pathologists, and analyzed using hierarchical clustering and principal component analysis to investigate protein expression profiles in the two variants of NSCLCs. Although Ullmann's approach showed that the two lung carcinomas subtypes can be classified using multiple antibody analysis, the method is time-consuming and very expensive.

While most studies have focussed on histotype specific proteins and their detection using immunohistochemistry, we explored the possibility of an automated method to deal with challenging H&E tissue patterns. In this paper, a robust computer vision approach to automatically classify AC and SC in real time using low cost H&E tissue images alone is introduced. The presented method includes a feature enhancement approach designed for H&E colour space, a tissue mor-

phological pattern extraction function and a machine learning algorithm. The proposed technique is fully automated, computationally efficient and has been demonstrated to robustly achieve high classification accuracy in our experiments.

Materials

Five micron tissue sections were taken from paraffin embedded samples, regions of interest defined by experienced pathologists and three tissue microarrays (TMA) generated using an MTA1 tissue arrayer (Beechers Instruments) with a core size of 0.6mm in diameter. In experiments, 369 tissue samples were collected in tissue microarray format, including 97 adenocarcinoma samples and 272 squamous carcinoma samples. The tumor regions were first delineated by an experienced histopathologist within which regions were selected for measurement. All the tissue slides were scanned using Aperio Scanscope CS2 (Aperio Technologies Inc. San Diego USA), at $\times 40$ objective magnification.

Methods

The proposed method includes (i) an image processing method (cw-HE) for generating distinctive hematoxylin and eosin staining patterns, reducing noisy information and producing discriminative image features, (ii) a feature measurement process to extract 37 densitometric and Haralick's [8] features from the H&E tissue sample images, and (iii) a boosting algorithm for pattern recognition.

Image Feature Enhancement Method of H&E Tissue Slides An image feature enhancement method, named cw-HE, is specifically designed for H&E images. The three major contributions of cw-HE are to produce distinctive H&E staining patterns, reduce noise and extract discriminative image data (see Fig. 1). As Hematoxylin induces blue staining of nuclei and Eosin induces the red/pink staining of cytoplasm, we applied 2D histogram equalization for image feature enhancement. It independently equalizes the two histogram distributions of the red channel and the blue channel in the RGB color space. Preliminary exploration showed that this performed better in separating these staining components than in HSL space (see Fig. 2). The images were then subjected to background subtraction to remove background noise, and the cw-HE enhanced images were subjected to subsequent feature extraction and analysis.

To further investigate the contribution of cw-HE, separate experiments on the TMA samples have been conducted utilizing the raw H&E grayscale images and the cw-HE processed images as presented in Table 2, showing that cw-HE greatly improves the classification accuracies for all machine learning methods.

Feature Extraction A total of 37 image features were extracted from each tissue core image including 22 Haralick's features [8] and 15 densitometric features (Table 1), utilizing Zeiss KS400 imaging system (Carl Zeiss, Oberkochen, Germany). Haralick's features are popularly adopted for texture classification, and

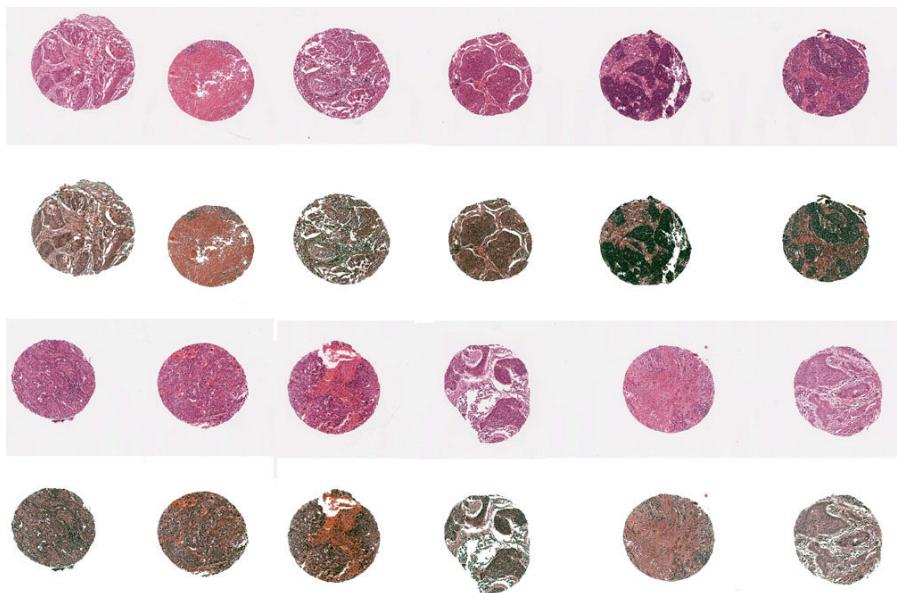


Fig. 1. Raw H&E tissue core images and enhanced cw-HE images, which produces discriminative image features for further pattern recognition

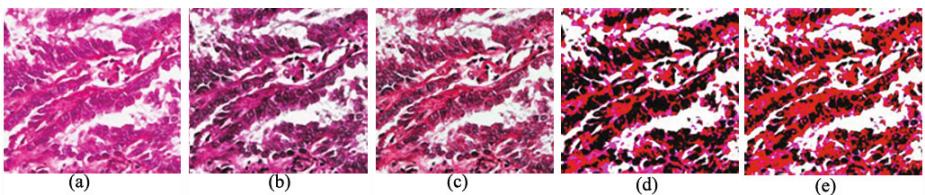


Fig. 2. Enhancing H&E staining patterns in RGB space performs better than in HSL space in separating nuclei and cytoplasm, evaluated with Otsu Clustering [11]: (a) input H&E image, (b) enhance H&E staining patterns in HSL color space, (c) enhance H&E staining patterns in RGB color space (d) apply Otsu clustering to the enhanced image in HSL space as shown in (b), (e) apply Otsu clustering to the enhanced image in RGB space as shown in (c). Comparing enhanced image in HSL space (d) and in RGB space (e), the tissue image enhanced in RGB space shows better separation with less nuclear misdetections.

densitometric features have been utilized for morphological classification in recent studies [3,19]. Each feature was then normalized to provide a value from 0 to 1 and used as the basis for tissue histology discrimination between NSCLC variants. Initially, this feature set was extracted from each of the red, green and blue components of the image. Initial studies however showed that the blue chan-

Table 1. Image features extracted

Texture	1-11	11 standard Haralick's features [8]
	12-22	11 mean values of individual Haralick features
Densitometric	23 24 25 26 27 28 29 30 31 32	number of bits of squares to store the densitometric values of the region sum sum of squares standard deviation skewness of squares kurtosis energy minimum intensity value maximum intensity value mean
Densitometric moments	33 34 35 36 37	x coordinate of the densitometric center of gravity of the region y coordinate of the densitometric center of gravity of the region length of the main axis of the ellipse with the same densitometric moment of inertia as the region length of the middle axis of the ellipse with the same densitometric moment of inertia as the region the angle of the main axis of the ellipse with the same densitometric moment of inertia as the region

nel had higher discriminative information in the classification of two NSCLCs tumor types than composite greyscale, red and green channels. Hence, the blue channel from each image was used in subsequent analyses and for developing the NSCLC classifier.

Machine Learning Algorithms For pattern discovery, we have extensively explored five popularly adopted machine learning approaches, including support vector machine [9], neural network [2], C4.5 decision tree [12], alternative decision tree [5] and two Boosting algorithms, including popularly adopted AdaBoost.M1 [6] and a low variance error boosting algorithm (cw-Boost) from the author's previous effort [17] in classification of high dimensional features but low numbers of instances such as gene expression datasets. For the machine learning methods, we adopted Weka [18] implementation of C4.5 decision tree (J48), alternative decision tree (ADTree), support vector machine (sequential minimal optimization (SMO)), neural network (Multilayer perceptron (MLP)) and AdaBoostM1. In evaluation, 10-fold cross validation [10] was used, and eight different classification algorithms are evaluated.

Results

Table 2 shows the 10-fold cross validation results of the analysis on the tissue microarray samples. The table shows that classification accuracy was improved across all classification methods when the cw-HE image enhancement methodology was applied, by comparison to the raw H&E image data. A comparison of classification methods shows that AdaBoostM1(ADT) and cw-Boost(ADT) performed the best with high accuracy of 92.14% and 91.06% respectively.

Table 2. 10-fold Cross Validation Results on 369 Tissue Samples of TMAs: the method (cw-HE + AdaB(ADT)) achieves the highest accuracy 92.14%. C4.5: C4.5 decision tree; ADT: alternative decision tree; SVM: support vector machine; NN: neural network; AdaB(C4.5) = use AdaBoostM1 as the ensemble machine learning algorithm and C4.5 decision tree as the base learner and here, we specify 110 base classifiers to build. Similarly, cwB(C4.5)= use cw-Boost as the ensemble and C4.5 decision tree as the base learner. Same definition is applied to AdaB(ADT) and cwB(ADTree).

	C4.5	ADT	SVM	NN	AdaB(C4.5)	AdaB(ADT)	cwB(C4.5)	cwB(ADT)
H&E	76.96	81.57	81.57	82.38	84.55	84.28	85.37	86.18
cw-HE	83.73	85.91	87.26	86.99	89.43	92.14	89.97	91.06

Table 3. Detailed Classification Accuracy By Cancer Types (AC=adenocarcinoma; SC=squamous carcinoma) of the best-performed classifiers for TMA samples samples using 10-fold cross validation.

best-performed classifier	Class	TP	FP	Precision	Recall	F-Measure	ROC
	SC	0.952	0.165	0.942	0.952	0.947	0.927
	AC	0.835	0.048	0.862	0.835	0.848	0.96

Discussion

The simple morphological distinction of small cell and non-small cell is no longer sufficient in describing clinical relevance of lung cancer and in NSCLC the histological subdivision into SC and non-SC tumours is extremely important role in the selection of lung cancer patients for therapy. However recent studies have shown poor reproducibility in the diagnosis of NSCLC SC tumours by both non-experts and experts [7] and it is recognized that new approaches are necessary [14]. For this reason, other methods are being explored such as histochemical stains (mucicarmine, PAS diastase) immunohistochemistry of histotype specific proteins (thyroid transcription factor-1, surfactant apoprotein, p63, cytokeratins 5/6) and molecular markers (e.g. HAS-MIR-205 expression) to better distinguish lesions [1]. For example, Ullmann et al. [16] presented a method to classify adenocarcinomas and squamous carcinomas using immunohistochemistry of a panel of 86 proteins, manual scoring by pathologists and hierarchical clustering and principle components analysis. While exploring and utilizing these biomarkers clearly has value, it is time consuming and adds to the overall cost of the diagnostic evaluation. A simpler solution would be to develop a more objective reproducible classification method of basic H&E samples, a method that that exceeds the performance of the human eye and improves classification of squamous and non-squamous tumours.

In this paper, we have developed a new fully-automated machine vision approach, aimed at classifying AC and SC in real time using conventional H&E tissue samples alone. There is an enormous body of literature that spans several decades demonstrating the value of morphometry and image analysis in tissue

histology. This study has demonstrated for the first time the value of tissue pattern measurement in distinguishing SC and non-SC histotypes. The methods used can be applied to standard H&E samples and does not rely on the delineation of individual cells or nuclei.

This would facilitate diagnostic classification and for the first time integrate computer vision and routine decision making in lung cancer. There are other advantages of the presented technique. In tissue microarray studies and the search for new biomarkers of prognosis, it is important to continually monitor the underlying histopathology of each core on subsequent sections from the TMA block. This is a very time consuming process when carried out visually by a pathologist. Automated analytical methods such as described here could be used to ensure that the underlying histopathological classification of a new section does not differ from the assignment in the original TMA map.

In conclusion therefore, this work represents a starting point in providing an objective basis to the classification of squamous and non-squamous in NSCLC, with primary advantages in the selection of patients for therapy now, but also as a research tool for improving the identification of other tumour sub-types that will benefit from selective and individualized therapy in the future.

References

1. Argiris, A., Gadgeel, S.M., Dacic, S.: Subdividing nsclc: Reflections on the past, present, and future of lung cancer therap. Oncology 23, 1–4 (2009)
2. Bishop, C.M.: Neural Networks for Pattern Recognition. Oxford University Press, Inc., New York (1995)
3. Chapman, J., Miller, N., Lickley, H., Qian, J., Christens-Barry, W., Fu, Y., Yuan, Y., Axelrod, D.: Ductal carcinoma in situ of the breast (dcis) with heterogeneity of nuclear grade: prognostic effects of quantitative nuclear assessment. BMC Cancer 7(1), 174 (2007)
4. Dubey, S., Powell, C.A.: Update in lung cancer 2008. Am. J. Respir. Crit. Care Med. 179(10), 860–868 (2009)
5. Freund, Y., Mason, L.: The alternating decision tree learning algorithm. In: Proc. 16th International Conf. on Machine Learning, pp. 124–133. Morgan Kaufmann, San Francisco (1999)
6. Freund, Y., Schapire, R.E.: Experiments with a new boosting algorithm. In: In Proceedings of the Thirteenth International Conference on Machine Learning, pp. 148–156 (1996)
7. Grilley-Olson, J.E., Hayes, D.N., Qaqish, B.F., Moore, D.T., Socinski, M.A., Yin, X., Leslie Wilkerson, K.O., Travis, W.D., Funkhouser, W.K., et al.: Validation of inter-observer agreement in lung cancer assessment. Journal of Clinical Oncology 27, 15 (2009)
8. Haralick, R.M., Shanmugam, K., Dinstein: Textural features for image classification. IEEE Transactions on Systems, Man and Cybernetics 3(6), 610–621 (1973)
9. Keerthi, S.S., Shevade, S.K., Bhattacharyya, C., Murthy, K.R.K.: Improvements to platt's smo algorithm for svm classifier design. Neural Computation 13(3), 637–649 (2001)
10. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. In: IJCAI, pp. 1137–1145 (1995)

11. Otsu, N.: A threshold selection method from gray level histograms. *IEEE Trans. Systems, Man and Cybernetics* 9, 62–66 (1979); minimize inter class variance
12. Quinlan, R.J.: C4.5: Programs for Machine Learning. Morgan Kaufmann Series in Machine Learning. Morgan Kaufmann, San Francisco (1993)
13. Sandler, A., Gray, R., Perry, M.C., Brahmer, J., Schiller, J.H., Dowlati, A., Lilienbaum, R., Johnson, D.H.: Paclitaxel-carboplatin alone or with bevacizumab for non-small-cell lung cancer. *N. Engl. J. Med.* 355(24), 2542–2550 (2006)
14. Selvaggi, G.: Histologic subtype in nsclc: Does it matter? *Oncology* 23, 1–11 (2009)
15. American Cancer Society, <http://www.cancer.org/> (accessed June 9, 2009)
16. Ullmann, R., Morbini, P., Halbwedl, I., Bongiovanni, M., Gogg-Kammerer, M., Papotti, M., Gabor, S., Renner, H., Popper, H.H.: Protein expression profiles in adenocarcinomas and squamous cell carcinomas of the lung generated using tissue microarrays. *J. Pathol.* 203(3), 798–807 (2004)
17. Wang, C.-W., Hunter, A.: A low variance error boosting algorithm. *Applied Intelligence* (2009)
18. Witten, I.H., Frank, E.: Data Mining: Practical Machine Learning Tools and Techniques, 2nd edn. Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, San Francisco (2005)
19. Zapotoczny, P., Zielinska, M., Nita, Z.: Application of image analysis for the varietal classification of barley: Morphological features. *Journal of Cereal Science* 48(1), 104–110 (2008)

Hot Spots Conjecture and Its Application to Modeling Tubular Structures

Moo K. Chung^{1,2,3,4}, Seongho Seo⁴, Nagesh Adluru², and Houri K. Vorperian³

¹ Department of Biostatistics and Medical Informatics

² Waisman Laboratory for Brain Imaging and Behavior

³ Vocal Tract Development Laboratory, Waisman Center
University of Wisconsin, Madison, WI 53706, USA

⁴ Department of Brain and Cognitive Sciences

Seoul National University, Korea

mkchung@wisc.edu

Abstract. The second eigenfunction of the Laplace-Beltrami operator follows the pattern of the overall shape of an object. This geometric property is well known and used for various applications including mesh processing, feature extraction, manifold learning, data embedding and the minimum linear arrangement problem. Surprisingly, this geometric property has not been mathematically formulated yet. This problem is directly related to the somewhat obscure *hot spots conjecture* in differential geometry. The aim of the paper is to raise the awareness of this nontrivial issue and formulate the problem more concretely. As an application, we show how the second eigenfunction alone can be used for complex shape modeling of tubular structures such as the human mandible.

1 Introduction

The second eigenfunction of the Laplace-Beltrami operator is drawing significant attention in recent years mainly as a tool for extracting shape features in high dimensional data [2,8,10]. The gradient of eigenfunction tends to follow the pattern of the overall shape of data and has been used to establish the intrinsic coordinate system of the data. This geometric property has been well known and often been used in computer vision and medical imaging applications. The second eigenfunction of the graph Laplacian was used to construct the Laplacian eigenmaps for low dimensional embedding [2]. The critical points of the second eigenfunction were used as anatomical landmarks for piecewise registration of colon surfaces [8]. The Reeb graph of the second eigenfunction was used in characterizing hippocampus shape [10].

All these studies rely on the geometric property of the second eigenfunction and somehow captures the overall shape of data. However, this property has not been mathematically formulated precisely. In fact, it is related to an obscure conjecture called the hot spots conjecture in differential geometry [1]. In this

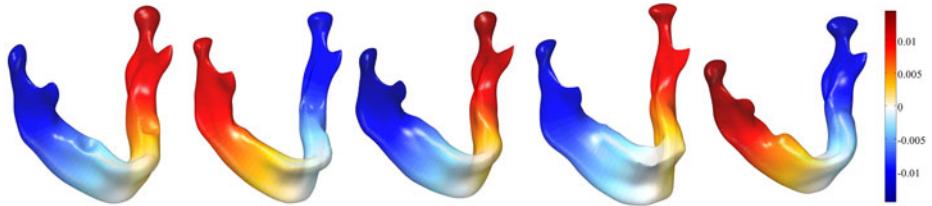


Fig. 1. The second eigenfunction ψ_1 for different mandible surfaces. The second eigenfunction for an elongated closed object is a smooth monotonic function increasing from one tip of surface to the other tip of the surface.

paper, we will explore the issue in detail and formulate the geometric property more precisely, which will be illustrated with examples and a proof for simple shape.

2 The Second Laplace-Beltrami Eigenfunction

Second Eigenfunction. The eigenfunctions of the Laplace-Beltrami operator has been used in many different contexts in image analysis [8,9,10]. Eigenfunctions ψ_j of Laplace-Beltrami operator Δ in \mathcal{M} satisfy $\Delta\psi_j = \lambda_j\psi_j$. The eigenfunctions form an orthonormal basis in \mathcal{M} . We can order the eigenfunctions $\psi_0, \psi_1, \psi_2, \dots$ corresponding to the increasing order of eigenvalues $0 = \lambda_0 \leq \lambda_1 \leq \lambda_2 \leq \dots$. Other than $\psi_0 = 1/\sqrt{\mu(\mathcal{M})}$, the close form expression for other eigenfunctions are unknown. However, using the cotan discretization for the Laplace-Beltrami operator [3,9], we can obtain the eigenfunctions numerically. The MATLAB code is available at <http://brainimaging.waisman.wisc.edu/~chung/lb>.

The critical point of the second eigenfunction ψ_1 usually occur at the two extremes of an elongated object (Figure 1). So the gradient of the second eigenfunction follows the shape of elongated objects. In computer vision literature, this *monotonicity property* was observed but without any mathematical justification [2,10]. Many treated it as a proven fact although the underlying conjecture has yet to be proved [1].

Conjecture 1. (Rauch's hot spots conjecture) [1] Let \mathcal{M} be an open connected bounded subset. Let $f(\sigma, p)$ be the solution of heat equation

$$\frac{\partial f}{\partial \sigma} = \Delta f \quad (1)$$

with the initial condition $f(0, p) = g(p)$ and the Neumann boundary condition $\frac{\partial f}{\partial n}(\sigma, p) = 0$ on the boundary $\partial\mathcal{M}$. Then for *most* initial conditions, if p_{hot} is a point at which the function $f(\cdot, p)$ attains its maximum (hot spot), then the distance from p_{hot} to $\partial\mathcal{M}$ tends to zero as $\sigma \rightarrow \infty$ [1].

We can also claim a similar statement for minimum (cold spots) as well. Conjecture 1 basically implies that the hot and cold spots move away from the

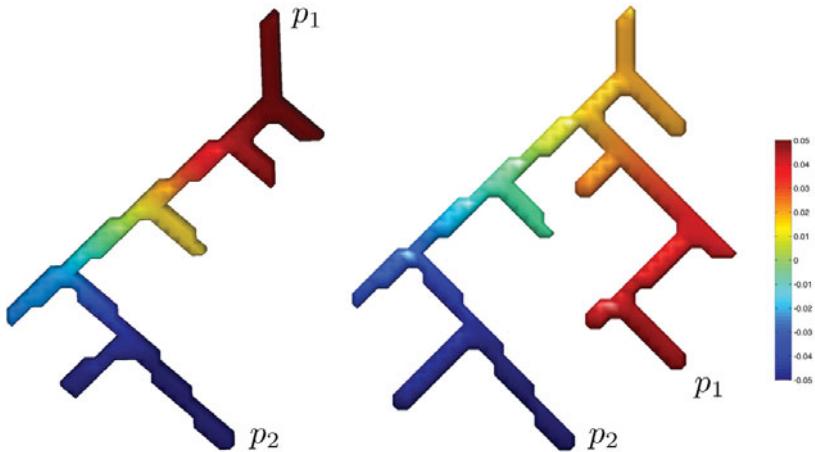


Fig. 2. The second eigenfunction ψ_1 of branching tubular structures. The maximum (p_1) and the minimum (p_2) of ψ_1 always occur at the points of maximum geodesic distance.

origin toward the boundary as the diffusion continues. Note that the solution to heat equation (1) is given by heat kernel expansion [9]:

$$K_\sigma * g(p) = \sum_{j=0}^{\infty} e^{-\lambda_j \sigma} \beta_j \psi_j(p), \quad (2)$$

where $\beta_j = \langle \psi_j, g \rangle$ are Fourier coefficients. Since $\lambda_0 = 0$ and $\psi_0 = 1/\sqrt{\mu(\mathcal{M})}$, we have

$$K_\sigma * g(p) = \frac{\int_{\mathcal{M}} g(p) d\mu(p)}{\mu(\mathcal{M})} + \beta_1 e^{-\lambda_1 \sigma} \psi_1(p) + R(\sigma, p), \quad (3)$$

where the first term is the average signal and the remainder R goes to zero faster than $e^{-\lambda_1 \sigma}$ as $\sigma \rightarrow \infty$ [1]. Therefore, the behavior of the propagation of the hot spots is basically governed by the eigenfunction ψ_1 .

What will happen if there is no boundary? In the case of closed manifold with no boundary, the Neumann boundary condition simply disappears so the direct application of the hot spots conjecture is not valid. Since ψ_1 asymptotically behaves like the heat equilibrium state, hot and cold spots cannot possibly be located in close proximity. Thus, we propose the following conjecture.

Conjecture 2. For a closed and sufficiently smooth simply connected surface \mathcal{M} with no boundary, the geodesic distance d between any two points p and q is bounded by

$$d(p, q) \leq d(p_{min}, p_{max}).$$

Conjecture 2 implies that the hot and cold spots give the maximum possible geodesic distance among all possible pairs of points and define the direction

of elongation of data. Figure 2 illustrates Conjecture 2. For any complicated branching binary tree structures like Figure 2, the hot and cold spots occur at the two extreme points along the longest geodesic path. The hot spots conjecture basically dictates that it is possible to find the maximum possible geodesic path by simply finding the critical points in the second eigenfunction. Conjecture 2 can be applicable not only to differentiable manifolds but to graphs and surface meshes as well.

3 Hot Spots Conjecture Applied to Fiedler's Vector

Surface meshes can be considered as graphs. The connection between the eigenfunctions of continuous and discrete Laplacians has been well established by many authors [6,11]. Many properties of eigenfunctions of the Laplace-Beltrami operator have discrete analogues. The second eigenfunction of the discrete graph Laplacian is called the Fiedler vector and it has been studied in connection to the graph and mesh processing, manifold learning and the minimum linear arrangement problem [5] Let $G = \{V, E\}$ be the graph with the vertex set V and the edge set E . G is the discrete approximation of the underlying continuous manifold \mathcal{M} . We will simply index the node set as $V = \{1, 2, \dots, n\}$. If two nodes i and j form an edge, we denote it as $i \sim j$. Various forms of graph Laplacian have been proposed but many graph or discrete Laplacian $L = (l_{ij})$ is a real symmetric matrix of the form

$$l_{ij} = \begin{cases} -w_{ij}, & i \sim j \\ \sum_{i \neq j} w_{ij}, & i = j \\ 0, & \text{otherwise} \end{cases}$$

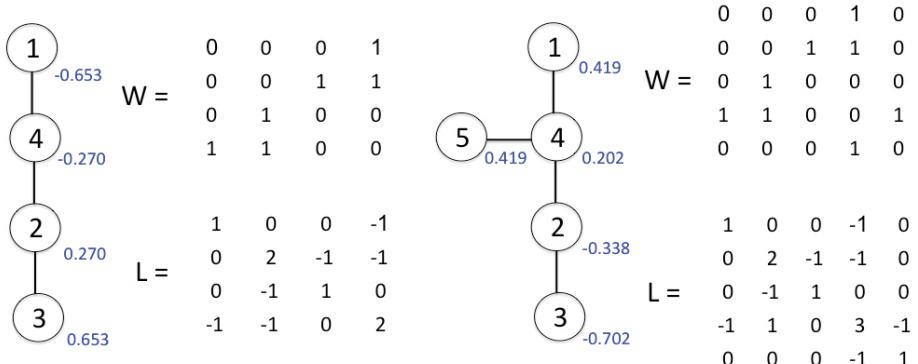


Fig. 3. A graph, the weights W and the graph Laplacian L . The weights are simply the adjacency matrix. The second eigenfunction ψ_1 value is displayed in blue. (a) This example is given in [7]. The maximum geodesic distance is obtained between the nodes 1 and 3, which are also hot and cold spots. (b) In this example, there are two hot spots 1 and 5 which correspond to two maximal geodesic paths 1-4-2-3 and 5-4-2-3.

for some edge weight w_{ij} . The graph Laplacian L can be decomposed as $L = D - W$, where $D = (d_{ij})$ is the diagonal matrix with $d_{ii} = \sum_{j=1}^n w_{ij}$ and $W = (w_{ij})$. For a vector $\mathbf{f} = (f_1, \dots, f_n)'$ observed at the n nodes, the discrete analogue of the Dirichlet energy is given by

$$\mathcal{E}(f) = \mathbf{f}' L \mathbf{f} = \sum_{i,j=1}^n w_{ij} (f_i - f_j)^2 = \sum_{i \sim j} w_{ij} (f_i - f_j)^2. \quad (4)$$

The Fiedler's vector $\mathbf{f} = (f_1, \dots, f_n)'$ is obtained as the minimizer of the quadratic polynomial $\psi_1 = \arg \min_f \mathcal{E}(\mathbf{f})$ subject to the quadratic constraint $\|\mathbf{f}\|^2 = \sum_i f_i^2 = 1$. The Dirichlet energy measures the smoothness of \mathbf{f} so the second eigenfunction should be the smoothest possible map among all possible functions. Since ψ_1 is required to be orthonormal with ψ_0 , we also have an additional constraint $\sum_i f_i = 0$. Therefore, ψ_1 is positive on half of \mathcal{M} and negative on the other half. However, it still does not explicitly tell us that a smooth function has to be monotonically changing from one end to the other.

The constraints force ψ_1 to have at least two differing *sign domains* in which ψ_1 has one sign. The *nodal set* of eigenfunctions ψ_i is defined as the zero level set $\psi_i(p) = 0$. Then Courant's nodal line theorem states that the nodal set of the i -th eigenfunction ψ_{i-1} divide the manifold into no more than i sign domains [4,6,11]. Hence, the second eigenfunction must have exactly 2 disjoint sign domains. At the positive sign domain, we have the global maximum and at the negative sign domain, we have the global minimum. This is illustrated in Figure 3 and 4. Although it is difficult to prove the general statement, the conjecture can be proven for specific cases. Here we provide a first heuristic proof for a path, which is a graph with maximal degree 2 and without a cycle.

Tightness. For a function \mathbf{f} defined on the vertex set V , let G_s^- be the subgraph induced by the vertex set $V_s^- = \{i \in V | f_i < s\}$. Similarly, let G_s^+ be the subgraph induced by the vertex set $V_s^+ = \{i \in V | f_i > s\}$. For any s , if G_s^- and G_s^+ are either connected or empty, then \mathbf{f} is tight [11]. The concept of tightness is crucial in proving the statement. When $s = 0$, G_0^+ and G_0^- are sign graphs. If we relax the condition so that G_s^+ contains nodes satisfying $f_i \geq s$, we have *weak* sign graphs. The second eigenfunction on a graph with maximal degree 2 (either cycle or path) is tight [11]. Figure 4 shows an example of a path with 11 nodes. Among three candidates for the second eigenfunction, (a) and (b) are not tight while (c) is. Note that the candidate function (a) has two disjoint components when thresholded at $s = 0.5$ so it cannot be tight. In order to be tight, the second eigenfunction cannot have a positive minimum or a negative maximum at the interior vertex in the graph [6]. This implies that the second eigenfunction must decrease monotonically from the positive to negative sign domains as shown in (c) and have the critical points at the two end points. This has to be true for a general case as well.

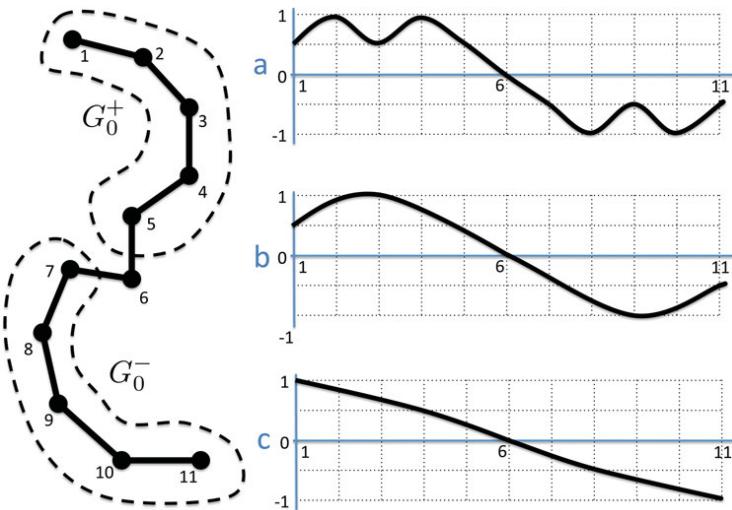


Fig. 4. A path with positive (G_0^+) and negative (G_0^-) sign domains. Among many possible candidate functions, (a) and (b) are not tight so they can't be the second eigenfunctions.

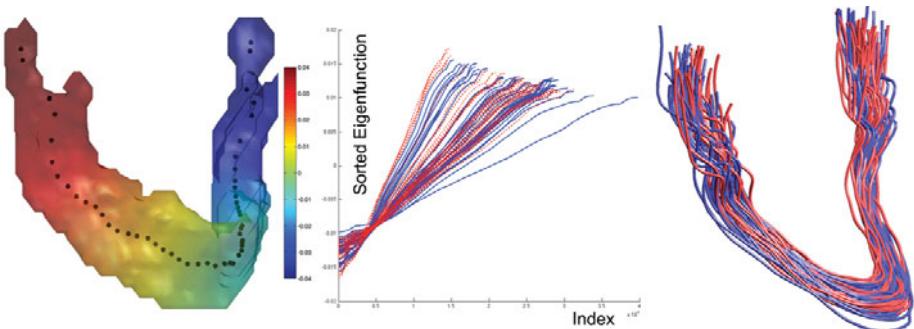


Fig. 5. Left: The centroids of level contours of the 2nd eigenfunction. Middle: The sorted second eigenfunctions (female = red, male = blue). Right: The centerlines of all 76 subjects showing the proper alignment.

4 Application: Mandible Growth Modeling

The CT imaging data set consists of 76 mandibles (40 males and 36 females). The age distribution was 11.33 ± 5.60 years for the females, and 9.54 ± 5.71 years for the males. The image acquisition and processing details for acquiring the mandible surface meshes are given in [9].

Features. In most literature dealing with the eigenfunctions of the Laplace-Beltrami operator, the whole spectrum of eigenvalues or eigenfunctions were used for shape analysis [9,10]. Here, we show how the second eigenfunction alone can be used as the shape feature. Once we obtained the second eigenfunctions for all subjects, we sorted them in increasing order (Figure 5). The sorted eigenfunctions are almost straight lines. A bigger mandible (more indices) exhibits a less steeper slope. Therefore, the rate of increase of the sorted eigenfunction (slope of linear fit) can be used for characterizing subject anatomical variability. *General Linear Models.* We examined how the rate of increase can be used in characterizing the growth of the mandible. We used the general linear model (GLM) of the form $\text{feature} = \beta_0 + \beta_1 \text{gender} + \beta_2 \text{age} + \epsilon$. The parameters are estimated using the least squares method and the statistical significance is determined using the F -statistic. There is weakly significant gender difference (β_1) (p -value = 0.08) and highly significant age effect (β_2) (p -value < 10^{-7}) for the rate of increase. We conclude that the mandible size grows at a much faster rate for males than females.

Acknowledgement. This work was funded by grants from NIH R01 DC6282 and P-30 HD03352. Also, WCU grant to the Department of Brain and Cognitive Sciences, Seoul National University. We thank Lindell R. Gentry, Mike S. Schimek, Katelyn J. Kassulke and Reid B. Durtschi for assistance with image acquisition and segmentation.

References

1. Banuelos, R., Burdzy, K.: On the Hot Spots Conjecture of J Rauch. *Journal of Functional Analysis* 164, 1–33 (1999)
2. Belkin, M., Niyogi, P.: Laplacian eigenmaps and spectral techniques for embedding and clustering. In: *Advances in Neural Information Processing Systems*, vol. 1, pp. 585–592 (2002)
3. Chung, M.K., Taylor, J.: Diffusion smoothing on brain surface via finite element method. In: *Proceedings of IEEE International Symposium on Biomedical Imaging (ISBI)*, vol. 1, pp. 432–435 (2004)
4. Courant, R., Hilbert, D.: *Methods of Mathematical Physics*, english edn. Interscience, New York (1953)
5. Fiedler, M.: Algebraic connectivity of graphs. *Czechoslovak Mathematical Journal* 23, 298–305 (1973)
6. Gladwell, G.M.L., Zhu, H.: Courant’s nodal line theorem and its discrete counterparts. *The Quarterly Journal of Mechanics and Applied Mathematics* 55, 1–15 (2002)
7. Hall, K.M.: An r-dimensional quadratic placement algorithm. *Management Science* 17, 219–229 (1970)
8. Lai, Z., Hu, J., Liu, C., Taimouri, V., Pai, D., Zhu, J., Xu, J., Hua, J.: Intra-patient supine-prone colon registration in CT colonography using shape spectrum. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) *MICCAI 2010. LNCS*, vol. 6361, pp. 332–339. Springer, Heidelberg (2010)

9. Seo, S., Chung, M.K., Vorperian, H.K.: Heat kernel smoothing using laplace-beltrami eigenfunctions. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6363, pp. 505–512. Springer, Heidelberg (2010)
10. Shi, Y., Lai, R., Krishna, S., Sicotte, N., Dinov, I., Toga, A.W.: Anisotropic Laplace-Beltrami eigenmaps: Bridging Reeb graphs and skeletons. In: Proceedings of Mathematical Methods in Biomedical Image Analysis (MMBIA), pp. 1–7 (2008)
11. Tlusty, T.: A relation between the multiplicity of the second eigenvalue of a graph laplacian, courants nodal line theorem and the substantial dimension of tight polyhedral surfaces. Electrnoic Journal of Linear Algebra 16, 315–324 (2007)

Fuzzy Statistical Unsupervised Learning Based Total Lesion Metabolic Activity Estimation in Positron Emission Tomography Images

Jose George¹, Kathleen Vunckx², Sabine Tejpar³, Christophe M. Deroose²,
Johan Nuyts², Dirk Loeckx¹, and Paul Suetens^{1,4}

¹ Center for Processing Speech and Images, Department of Electrical Engineering

² Department of Nuclear Medicine, ³ Department of Gastroenterology

^{1,2,3} Katholieke Universiteit Leuven, Belgium

⁴ IBBT-K.U.Leuven Future Health Department

Abstract. Accurate tumor lesion activity estimation is critical for tumor staging and follow up studies. Positron emission tomography (PET) successfully images and quantifies the lesion metabolic activity. Recently, PET images were modeled as a fuzzy Gaussian mixture to delineate tumor lesions accurately. Nonetheless, on the course of accurate delineation, chances are high to potentially end up with activity underestimation, due to the limited PET resolution, the reconstruction images suffer from partial volume effects (PVE). In this work, we propose a statistical lesion activity computation (SLAC) approach to robustly estimate the total lesion activity (TLA) directly from the modeled Gaussian partial volume mixtures. To evaluate the proposed method, synthetic lesions were simulated and reconstructed. TLA was estimated from 3 state-of-the-art PET delineation schemes for comparison. All schemes were evaluated with reference to the ground truth knowledge. The experimental results convey that the SLAC is robust enough for clinical use.

Keywords: Positron emission tomography, tumor activity estimation, finite mixture models, Gaussian distribution, partial volume modeling, linear combination (LC) of random variables.

1 Introduction

Over the past decade, PET imaging has been accentuating its role in the field of oncology [1]. For a wide range of clinical practices, spanning from radiotherapy treatment to tumor staging and patient follow up evaluations, ¹⁸F-FDG PET has established itself to be an indispensable tool [2]. In recent clinical studies [3,4,5], PET volume delineation and the subsequent quantification of tumor lesion activity drew immense interest among oncologists. In some cases, PET imaging detects changes in the tumor metabolic activity even hours after drug administration, and may therefore be capable of predicting therapy response well before the current gold standard computed tomography (CT) [5]. Radiotherapy, on the other hand, needed better tumor localization so as to selectively annihilate the

proliferating cancerous lesions. For tumor staging and therapy follow up evaluations, better reproducibility and repeatability in lesion activity quantification is considered vital rather than precision in tumor position estimation. In this paper, we focus on robust estimation of the lesion metabolic activity rather than tumor localization.

1.1 Related Works

Caillo et al. proposed a fuzzy Gaussian mixture model using stochastic expectation maximization (SEM) framework with a stationary Markov random field (MRF) to model the partial pixels for remote sensing applications in [6], [7]. Van Leemput et al. independently proposed a similar unified approach to model partial voxels in magnetic resonance images using Gaussian mixture based expectation maximization (EM) algorithm with a MRF [8]. Hatt et al. applied the fuzzy SEM approach to model partial volumes in the PET image [9]. This framework was further extended for modeling heterogeneous tumors in [10]. Here, we use the fuzzy Gaussian SEM algorithm in [6], [7], [9] to model the PET image and propose a statistical method for the direct estimation of TLA, modeling the partial volume voxels as a linear combination of tumor and background voxels.

2 Materials and Methods

2.1 Total Lesion Metabolic Activity and Statistical TLA

Total lesion activity (TLA) can be defined as the totality of metabolic activity integrated over the whole lesion. Here we derive the mathematical interpretation of TLA and its statistical counterpart as follows. Let the observed PET image and the hidden segmentation map be realizations $\mathbf{y} = \{y_s\}_{s \in \mathcal{S}}$ and $\mathbf{x} = \{x_s\}_{s \in \mathcal{S}}$ of the random fields $\mathcal{Y} = \{\mathcal{Y}_s\}_{s \in \mathcal{S}}$ and $\mathcal{X} = \{\mathcal{X}_s\}_{s \in \mathcal{S}}$ respectively, where $\mathcal{S} = \{1, \dots, N\}$ is the set of voxels. Integrating the activity, y_s over the whole \mathcal{S} , the total PET activity (TPA), constituting contributions from the lesion (TLA) and the background activities, can be mathematically defined as

$$\text{TPA} := \int y_s ds = \int \xi h(\xi) d\xi \equiv N \int \xi f(\xi) d\xi = N \times E(\mathcal{Y}_s) \quad (1)$$

The observed histogram $h(\xi)$ containing the frequency of occurrence for each ξ (see the actual histogram in black shade in Fig. 1) can be related to the density $f(\xi)$ defining the distribution of $\mathcal{Y}_s = \xi$. $E(\mathcal{Y}_s)$ is the associated expectation.

Let $\mathbf{c} = \{c_k\}_{k \in \mathcal{K}}$, where $\mathcal{K} = \{1, \dots, Q\}$, be the Q class labels associated with the segmentation map. Then \mathcal{Y}_s can be modeled as a finite mixture of conditional densities as shown in Fig. 1 such that $f(\xi) = \sum_k \gamma_k f(\xi | c_k)$, where γ_k stands for the mixing probabilities $P(\mathcal{X}_s = c_k)$ and $f(\xi | c_k)$ denotes the density defining the distribution of $\mathcal{Y}_s = \xi$ conditional to $\mathcal{X}_s = c_k$. If $E(\mathcal{Y}_s | c_k)$ represents the conditional expectation of \mathcal{Y}_s given $\mathcal{X}_s = c_k$, then the TPA can be modified as

$$\text{TPA} := N \sum_k \gamma_k \left(\int \xi f(\xi | c_k) d\xi \right) = N \sum_k \gamma_k E(\mathcal{Y}_s | c_k) \quad (2)$$

To model \mathcal{Y}_s with the PVE, let the noise associated with the tumor and the background be modeled by 2 Gaussian random variables \mathcal{Y}_1 and \mathcal{Y}_0 with the density $\mathcal{N}(\mu_1, \sigma_1^2)$ ($f(\xi | \mathcal{X}_s = 1)$) and $\mathcal{N}(\mu_0, \sigma_0^2)$ ($f(\xi | \mathcal{X}_s = 0)$) defining the distributions of \mathcal{Y}_s conditional to $\mathcal{X}_s = 1$ and $\mathcal{X}_s = 0$ respectively. Then the partial volume activities can be modeled by linear combinations of these independent random variables as $\mathcal{Y}_s = (1 - \epsilon)\mathcal{Y}_0 + \epsilon\mathcal{Y}_1$, for $\mathcal{X}_s = \epsilon \in]0, 1[$. \mathcal{Y}_s is again Gaussian distributed with density $\mathcal{N}(\mu_\epsilon, \sigma_\epsilon^2)$ ($f(\xi | \mathcal{X}_s = \epsilon)$) conditional to $\mathcal{X}_s = \epsilon$, such that $\mu_\epsilon = (1 - \epsilon)\mu_0 + \epsilon\mu_1$ and $\sigma_\epsilon^2 = (1 - \epsilon)^2\sigma_0^2 + \epsilon^2\sigma_1^2$. Let M fuzzy LCs be used to model the partial volume mixtures, then $\mu_{\epsilon_k} = (1 - \epsilon_k)\mu_0 + \epsilon_k\mu_1$ with $\epsilon_k = k/(M + 1)$, $k = 1, \dots, M$. Therefore, for the tumor and the background, $\mu_{\epsilon_{M+1}} = \mu_1$ and $\mu_{\epsilon_0} = \mu_0$ respectively, with $\epsilon_0 = 0$ and $\epsilon_{M+1} = 1$. Here $\mathbf{c} = \{0, 1, \epsilon_k\}$, $Q = M + 2$. In eq. 2, out of the total activity in a fuzzy mixture class F_{ϵ_k} in $E(\mathcal{Y}_s | \epsilon_k)$, only $\epsilon_k\mu_1$ originates from the tumor lesion. Adding up the contributions from 1 hard class and M fuzzy classes representing tumor and partial volume voxels respectively, we propose that, the statistical TLA can be estimated from TPA as

$$\text{TPA} := N \sum_k \gamma_k E(\mathcal{Y}_s | \epsilon_k) = N \sum_k \gamma_k \mu_{\epsilon_k} \implies \text{TLA} := N \sum_{k=1}^{M+1} \gamma_k \epsilon_k \mu_1 \quad (3)$$

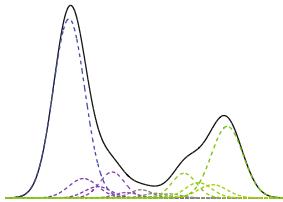


Fig. 1. Actual histogram (in black) modeled as a finite mixture of 2 hard classes (tumor in green and background in blue shade) and a mixture of fuzzy classes (in a combination of blue and green shades) representing partial volume voxels

2.2 Stochastic Expectation Maximization and TLA Estimation

The PET image is modeled as a fuzzy Gaussian mixture using SEM algorithm with locally adaptive prior update in each iteration as in [6], [7], [9]. To deal with initialization, we use 12 random starts and select the one which minimizes message length (MML) [11]. The steps involved in the proposed statistical lesion activity computation (SLAC) algorithm are listed in Algorithm 1.

2.3 State-of-the-Art PET Delineation Schemes

Three competent PET volume delineation methods were selected for this study, namely a stochastic (FLAB) [9], a gradient based (GDM) [12] and an adaptive threshold based (ATM) [13] method. These methods were selected as their performance has been reported to be superior to the threshold based region growing method, commonly used in clinical trials, over a wide variety of lesions.

FLAB uses the same fuzzy SEM algorithm as we use in SLAC, but then to delineate lesions accurately. In GDM, an adaptive bilateral filter first denoises

Algorithm 1. Fuzzy SEM TLA Estimation with Partial Volume Voxel Modeling

Inputs: Image voxels $\mathbf{y} = \{y_s\}_{s \in \mathcal{S}}$, # Random starts RS, Cut off err_{cutoff} , # Iterations T

Outputs: TLASLAC

Initialization:

```

 $k \leftarrow 0$ , MML  $\leftarrow \infty$ 
while  $k < RS$  do
     $k \leftarrow k + 1$ ,  $\widehat{\Theta}_k \leftarrow FCMClustering(\mathbf{y})$ , where  $\widehat{\Theta}_k = \{\mu_C, \sigma_C^2, \gamma_C\}$ ,  $C = \{0, 1, \mathcal{F}\}$ 
    ML  $\leftarrow MessageLength(\mathbf{y}, \widehat{\Theta}_k)$  computed as in [11]
    if  $ML < MML$  then
        MML  $\leftarrow ML$ ,  $\Theta_{opt} \leftarrow \widehat{\Theta}_k$ 
    end if
end while

```

Stochastic Expectation Maximization:

```

 $t \leftarrow 0$ , TLASLAC  $\leftarrow 0$ ,  $\Theta \leftarrow \Theta_{opt}$ 

```

```

while  $t < T$  do

```

```

repeat

```

```

 $t \leftarrow t + 1$ ,  $w_C^{(s)} \leftarrow \frac{\gamma_C^{(s)} f(y_s | C)}{\gamma_0^{(s)} f(y_s | 0) + \gamma_1^{(s)} f(y_s | 1) + (1 - \gamma_0^{(s)} - \gamma_1^{(s)}) \int_0^1 f(y_s | \theta) d\theta} \forall s \in \mathcal{S}$ 
 $\mathcal{Z}_s \leftarrow \arg \max_C w_C^{(s)}$ 
if  $\mathcal{Z}_s = \mathcal{F}$  then
     $\mathcal{Z}_s \leftarrow \arg \max_{\epsilon_k} f(y_s | \epsilon_k)$ , to allocate right  $F_{\epsilon_k}$  in  $\mathcal{F} = \{F_{\epsilon_1}, \dots, F_{\epsilon_M}\}$ 
end if
 $\gamma_C^{(s)} \leftarrow \frac{1}{Card(\mathcal{W}_s)} \sum_{r \in \mathcal{W}_s} \delta(\mathcal{Z}_r, C)$ , where  $\mathcal{W}_s$  is a  $w \times w \times w$  local support window
 $\mu_C \leftarrow \frac{\sum y_s \delta(\mathcal{Z}_s, C)}{\sum \delta(\mathcal{Z}_s, C)}$ ,  $\sigma_C^2 \leftarrow \frac{\sum (y_s - \mu_C)^2 \delta(\mathcal{Z}_s, C)}{\sum \delta(\mathcal{Z}_s, C)}$ ,  $\gamma_C = \frac{\sum \delta(\mathcal{Z}_s, C)}{N}$ 
 $\mu_{\epsilon_k} \leftarrow (1 - \epsilon_k)\mu_0 + \epsilon_k\mu_1$ ,  $\sigma_{\epsilon_k}^2 \leftarrow (1 - \epsilon_k)^2 \sigma_0^2 + \epsilon_k^2 \sigma_1^2$ 
 $TLA \leftarrow N \sum_k \gamma_k \epsilon_k \mu_1$ 
 $err \leftarrow |TLA - TLASLAC|$ ,  $TLASLAC \leftarrow TLA$ 

```

```

until  $err < err_{cutoff}$ 

```

```

return TLASLAC

```

```

end while

```

the image and next deblurs it with Landweber's deconvolution algorithm. The resulting gradient image is segmented using watersheds. Finally, the granularity of the watershed segments is made coarser using Ward's hierarchical clustering. ATM is optimized for uniform spherical objects: it computes the optimal threshold as a function of the sphere size, the object-to-background activity ratio and the spatial resolution of the scanner. This relation can be computed and/or determined from phantom measurements beforehand. In addition to the above mentioned automated methods, a 40% threshold based region growing method (T40) was evaluated.

2.4 Simulated Lesions

To evaluate the performance, two types of phantoms with hot lesions were simulated. First, a warm, water-filled cylinder (200mm diameter, 155mm long) with six hot spheres was simulated. The spheres were located in the central plane on

a circle with a diameter of 110mm. The diameters (ϕ) of the spheres were 37, 28, 22, 17, 13, and 10mm (see Fig. 2). The cylinder was filled with 7.4kBq/cc. The lesion-to-background ratio was set to 4 : 1. In order to accurately model the spheres, 0.5mm × 0.5mm × 0.5mm voxels were used during simulation.

Next, realistic FDG uptake values were assigned to the various organs and tissues of the NCAT phantom [14] based on a clinical FDG PET scan. 8 non spherical tumors were inserted in the liver, lungs, spleen and lymph node (see Fig. 2). In all tumors, the activity was set to 18.2kBq/cc. The activity in the liver, spleen, lungs and body was 6.3kBq/cc, 5.5kBq/cc, 0.9kBq/cc and 2.5kBq/cc respectively. The voxel size used to generate the phantom was 1mm × 1mm × 1mm.

For the hot spheres phantom, 30 noisy fully 3D acquisitions of 1min on an ECAT Exact HR+ were simulated using an analytical projector. The scanner resolution was modeled by a Gaussian with a full width half maximum (FWHM) of 5mm. Attenuation was modeled as well. The detector sampling was 2.25mm in transaxial direction and 2.425mm in axial direction. To evaluate the algorithms on more realistic dataset, 30 3min scans of the NCAT phantom were simulated using a Monte Carlo simulator (PET-SORTEO [15]) which models among others the spatially variant point spread function (PSF) of the ECAT Exact HR+ scanner. Attenuation and scatter were also modeled. During reconstruction of both datasets, the system PSF resolution was recovered by modeling as an isotropic Gaussian with 5mm FWHM (as done in clinical reconstructions).

For both phantom, the projection data were reconstructed using the maximum likelihood expectation maximization (MLEM) algorithm [16] with ordered subsets. As in clinical routine, 4 iterations over 16 subsets were performed. The reconstruction voxel size was set to 2mm × 2mm × 2mm. The images were post-smoothed with 5mm Gaussian FWHM.

2.5 Quantitative Evaluation

Different segmentation methods were evaluated by comparing the total lesion activity (TLA) estimated from delineated PET volumes against a resampled fuzzy ground truth knowledge (8mm³ voxels). Statistical TLA (SLAC) was estimated explicitly as proposed in eq. 3 and compared. For quantifying the error in TLA, δ TLA was computed as δ TLA = $\left(\frac{\text{TLA}_{\text{Estimator}} - \text{TLA}_{\text{True}}}{\text{TLA}_{\text{True}}} \right) \times 100$ in %. $\text{TLA}_{\text{Estimator}}$ stands for TLA estimated from a delineation method or the proposed direct approach. TLA_{True} stands for the actual TLA in the ground truth.

3 Experiments

Volume of interest containing the tumor lesion and background was selected. In SLAC, 10 fuzzy levels and a 3 × 3 × 3 support window (see Algorithm 1) was used. FLAB was iterated 50 times unless the cut-off condition, 0.1% parameter change, was reached. The noise was modeled as Gaussian and 3 fuzzy levels were used along with 2 hard levels for the analysis. To obtain the binary delineation map,

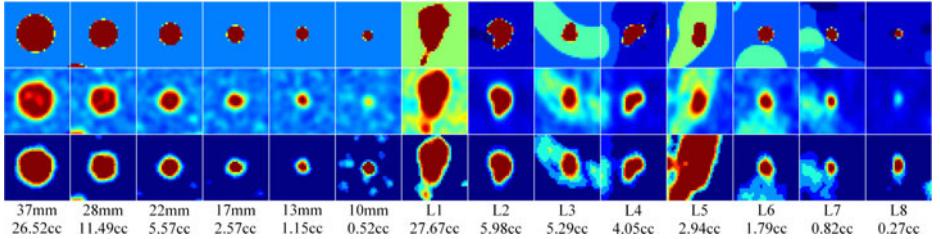


Fig. 2. Ground truth (1st row), MLEM reconstructed (2nd row) and fuzzy SEM modeled label (3rd row) image slices of simulated spherical (IEC Phantom - 1st to 6th column) and non spherical tumor lesions (NCAT Phantom - 7th to 14th column). Lesions L1 and L5 were in the liver whereas L2, L4, L7 and L8 were in the lungs. L3 was in the spleen and L6 in the lymph node.

the 2 fuzzy levels closest to the tumor class were taken as tumor and the other fuzzy level as background. For GDM, the denoising was iterated twice whereas the deconvolution ran over 30 iterations. ATM was set to run 100 iterations unless the cut-off condition, 0.001% relative threshold level (RTL) change, was reached. δ TLA was hence computed.

3.1 Simulated Lesions

Fig. 3 illustrates the median δ TLA for the proposed statistical TLA estimation method and the different segmentation methods, as a function of tumor lesions. We can see that δ TLA_{SLAC} estimated from the SLAC was found to be closer to the ideal scenario (δ TLA = 0), except for the smallest lesions (size < 1cc). Conventional delineation methods result in considerable under estimation of lesion metabolic activity (δ TLA_{FLAB}, δ TLA_{ATM}) for both sets of lesions, although deblurring helped GDM (δ TLA_{GDM}). Moreover, we can see that, even for the similar stochastic method, estimating activity from delineation is not the ideal way to go for ($|\delta$ TLA_{SLAC}| << | δ TLA_{FLAB}|). Even though, there were only background and partial voxels in the smallest spherical lesion ($\phi = 10mm$), our method tends to model it with tumor, background and partial classes (see Fig. 2) and hence it exhibits significant deviation from the true value (see Fig. 3(a)). For realistic tumors the deviation increases, since in most of the studied cases there were more than 2 dominant classes and hence modeling the system as 2 hard classes and a fuzzy LC of those hard classes tends to underperform. A typical example is the L5 lesion, where the 95% confidence interval (CI) (see Fig. 3(b)) was very big because 3 classes of data (L5 has liver, liver background and lung background) were modeled with just 2 hard classes (see Fig. 2).

4 Discussion and Conclusion

To our best knowledge, there are hardly any methods being developed taking into consideration the need to estimate actual metabolic activity. Moreover, due

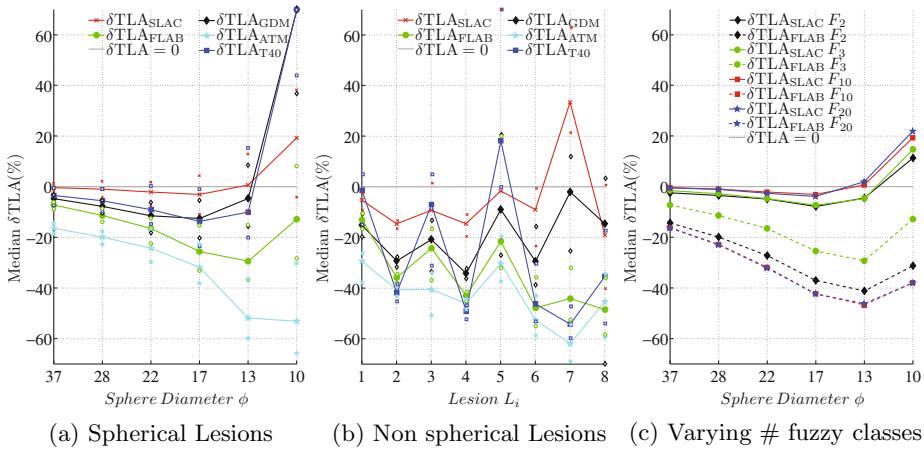


Fig. 3. Plots showing the median δTLA , along with 95% CI scatters, versus tumor lesions for the IEC phantom spherical lesions (in Fig. 3(a)) and the NCAT phantom non spherical lesions (in Fig. 3(b)). $\delta\text{TLA}_{\text{SLAC}}$ and $\delta\text{TLA}_{\text{FLAB}}$ (assigns top $\lfloor \frac{M}{2} \rfloor$ fuzzy classes as tumor) obtained with varying fuzzy levels for $3 \times 3 \times 3$ window, shown in (c).

to the limitation imposed by the PET acquisition system, any approach to delineate the actual tumor lesion will potentially end up underestimating the actual activity. This is in fact substantiated in the plots. We found that SLAC outperforms FLAB, although the methodology is basically the same. FLAB delineates the lesion, ignoring the estimated information about PVE, and then computes TLA; thereby loosing activity in the lower partial volume mixtures. SLAC, on the other hand, exploits the PVE model to have a better TLA estimate.

Further, we investigated the impact of higher support window and a larger number of fuzzy classes. While a window of $5 \times 5 \times 5$ was found to have little impact compared to $3 \times 3 \times 3$, an increase in fuzzy levels (> 10) avails better modeling of the partial volume voxels compared to 2, 3 or 5 fuzzy classes. Nevertheless, increasing the partial volume mixtures above 10 makes negligible difference (see Fig. 3(c)). The proposed SLAC algorithm is computationally fast (takes only a few seconds to process 30^3 voxels) and robust to initialization since it uses the SEM algorithm instead of more conventional EM algorithm. As a con, almost all methods including the SLAC showed increase in deviation (95% CI) with decrease in lesion size. This is because, there were more partial volume voxels in smaller lesions relative to the total tumor voxels. Noise also has a prominent role in this deviation.

The investigations made on the simulated spherical and non spherical lesions of varying size, shape and contrast indicate that our proposed method tends to give the best estimate compared to the state-of-the-art methods, especially when 2 classes (tumor and background) are present. The future work includes modeling with other probability distributions, multi-class data, clinical PET images, application to heterogeneous tumors and robustness study for varying reconstruction parameters.

Acknowledgments. The authors gratefully acknowledge the financial support by K.U.Leuven's Concerted Research Action GOA/11/006 and IWT Agency for Innovation by Science and Technology - Applied Biomedical Research (TBM) project 070717.

References

1. Weber, W.A., Figlin, R.: Monitoring cancer treatment with PET/CT: Does it make a difference? *JNM* 48, 36S–44S (2007)
2. Kelloff, G.J., Hoffman, J.M., Johnson, B., et al.: Progress and promise of FDG PET imaging for cancer patient management and oncologic drug development. *Clin. Cancer Res.* 11, 2785–2808 (2005)
3. Hatt, M., Visvikis, D., Albarghach, N., Tixier, F., Pradier, O., le Rest, C.C.: Prognostic value of ^{18}F -FDG PET image-based parameters in oesophageal cancer and impact of tumour delineation methodology. *EJNMMI* 38, 1191–1202 (2011)
4. Wahl, R.L., Jacene, H., Kasamon, Y., Lodge, M.A.: From RECIST to PERCIST: evolving considerations for PET response criteria in solid tumors. *JNM* 50, 122S–150S (2009)
5. Lucignani, G., Larson, S.M.: Doctor, What does my future hold? The prognostic values of FDG-PET in solid tumours. *EJNMMI* 37, 1032–1038 (2010)
6. Caillol, H., Hillion, A., Pieczynski, W.: Fuzzy Random Fields and Unsupervised Image Segmentation. *IEEE TGRS* 31, 801–810 (1993)
7. Caillol, H., Pieczynski, W., Hillion, A.: Estimation of fuzzy Gaussian mixture and unsupervised statistical image segmentation. *IEEE TIP* 6, 425–440 (1997)
8. Van Leemput, K., Maes, F., Vandermeulen, D., Suetens, P.: A unifying framework for partial volume segmentation of brain MR images. *IEEE TMI* 22, 105–119 (2003)
9. Hatt, M., le Rest, C.C., Turzo, A., Roux, C., Visvikis, D.: A fuzzy locally adaptive bayesian segmentation approach for volume determination in PET. *IEEE TMI* 28, 881–893 (2009)
10. Hatt, M., le Rest, C.C., Descourt, P., Dekker, A., De Ruyscher, D., Oellers, M., Lambin, P., Pradier, O., Visvikis, D.: Accurate automatic delineation of heterogeneous functional volumes in positron emission tomography for oncology applications. *Int. J. Radiation Oncology* 77, 301–308 (2010)
11. Figueiredo, M.A.T., Jain, A.K.: Unsupervised Learning of Finite Mixture Models. *IEEE PAMI* 24, 381–396 (2002)
12. Geets, X., Lee, J.A., Bol, A., Lonneux, M., Grégoire, V.: A gradient-based method for segmenting FDG-PET images: methodology and validation. *EJNMMI* 34, 1427–1438 (2007)
13. Van Dalen, J.A., Hoffmann, A.L., Dicken, V., Vogel, W.V., Wiering, B., Ruers, T.J., Karssemeijer, N., Oyen, W.J.G.: A novel iterative method for lesion delineation and volumetric quantification with FDG PET. *Nucl. Med. Communications* 28, 485–493 (2007)
14. Segars, W.P.: Development of a new dynamic NURBS-based cardiac-torso (NCAT) phantom. PhD Dissertation, The University of North Carolina (May 2001)
15. Reilhac, A., Lartizien, C., Costes, N., Sans, S., Comtat, C., Gunn, R.N., Evans, A.C.: PET-SORTEO: A monte carlo-based simulator with high count rate capabilities. *IEEE Trans. Nucl. Sci.* 51, 46–52 (2004)
16. Hudson, H.M., Larkin, R.S.: Accelerated image reconstruction using ordered subsets of projection data. *IEEE TMI* 13, 601–609 (1994)

Predicting Clinical Scores Using Semi-supervised Multimodal Relevance Vector Regression

Bo Cheng¹, Daoqiang Zhang^{1,2}, Songcan Chen¹, and Dinggang Shen²

¹ Dept. of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

² Dept. of Radiology and BRIC, University of North Carolina at Chapel Hill, NC 27599
{cb729, dqzhang, s.chen}@nuaa.edu.cn, dgshen@med.unc.edu

Abstract. We present a novel semi-supervised multimodal relevance vector regression (SM-RVR) method for predicting clinical scores of neurological diseases from multimodal brain images, to help evaluate pathological stage and predict future progression of diseases, e.g., Alzheimer’s diseases (AD). Different from most existing methods, we predict clinical scores from multimodal (imaging and biological) biomarkers, including MRI, FDG-PET, and CSF. Also, since mild cognitive impairment (MCI) subjects generally contain more noises in their clinical scores compared to AD and healthy control (HC) subjects, we use only their multimodal data (i.e., MRI, FDG-PET and CSF), not their clinical scores, to train a semi-supervised model for enhancing the estimation of clinical scores for AD and healthy control (HC). Experimental results on ADNI dataset validate the efficacy of the proposed method.

1 Introduction

Many pattern classification methods have been proposed for the diagnosis of Alzheimer’s disease (AD) or its prodromal stage, i.e., mild cognitive impairment (MCI). Recently, people have started to investigate applying the pattern regression methods for estimating the continuous clinical scores of subjects from their respective brain images [1,2,4]. This kind of study is important because it can help evaluate the pathological stage and predict the future progression of neurological diseases. It is known that many diseases present a continuous spectrum of structural and functional changes. For example, AD pathology is known to progress gradually over many years, sometimes starting decades before a final clinical stage [2]. Thus, pattern regression methods can be used to help estimate the continuous clinical scores to evaluate the disease stage of MCI or AD, rather than simple categorical classification.

Recent studies have also demonstrated that the biomarkers from different modalities can provide complementary information for diagnosis of AD [5,8,9]. Accordingly, two or more modalities of biomarkers have been combined for multimodal classification [3,6,7,14]. However, to the best of our knowledge, there exist few related works which combine two or more biomarkers from multimodal data for regression. Instead, most existing methods on estimating clinical scores use only single modality of data [1,2,4]. In this paper, we will employ multiple-kernel combination method to combine multimodality data, e.g., MRI, PET, and CSF, for multimodal regression.

On the other hand, at present, several works have adopted the supervised relevance vector machine regression (RVR) method to estimate continuous clinical scores [1,2,4]. It is known that the training of a supervised model often requires many well-labeled data in order to achieve a good performance. However, the clinical scores in cognitive tests such as Mini Mental State Examination (MMSE) and Alzheimer’s Disease Assessment Scale-Cognitive subtest (ADAS-Cog) are usually very noisy, especially for the MCI subjects who may or may not convert to AD within a period of follow-up time. To partially alleviate this problem, we will not exploit the clinical scores of the MCI subjects and instead use only their corresponding multimodal data (i.e., MRI, FDG-PET, and CSF) to help train a semi-supervised regression model. It is worth noting that similar idea, which treats MCI subjects as unlabeled data to train a semi-supervised *classification* model, has also been used for classification of AD [13,15]. However, to our knowledge, no previous studies have ever investigated the semi-supervised *regression* in AD research.

In this paper, we propose a semi-supervised multimodal relevance vector regression (SM-RVR) model to predict clinical scores based on both imaging and biological biomarkers. We further construct a graph Laplacian matrix based on the manifold learning theory [11] to select the most informative MCI subjects for better helping semi-supervised regression.

2 Method

In this section, we will first extend the standard relevance vector regression (RVR) method to the multimodal RVR (M-RVR), and then introduce our proposed semi-supervised multimodal RVR (SM-RVR) method.

2.1 Multimodal RVR (M-RVR)

We first briefly review the standard RVR algorithm. The main idea of RVR is summarized as follows. Specifically, RVR is a sparse kernel method formulated in a Bayesian framework [12]. Given a training set with its corresponding target values, such as $\{x_n, t_n\}_{n=1}^N$, RVR aims to find out the relationship between the input feature vector x_n and its corresponding target value t_n :

$$t_n = f(x_n, w) + \varepsilon_n \quad (1)$$

where ε_n is the measurement noise (assumed independent and following a zero-mean Gaussian distribution, $\varepsilon_n \sim N(0, \sigma^2)$), and $f(x_n, w)$ is a linear combination of basis functions $k(x, x_n)$ with the following form:

$$f(x, w) = \sum_{n=1}^N w_n k(x, x_n) + w_0 \quad (2)$$

Where $w = (w_0, w_1, \dots, w_N)^T$ is a weight vector, $K_{N \times (N+1)}$ is the ‘design’ matrix with $K_{ij} = k(x_i, x_j)$, $i, j = 1, \dots, N$, and $k(x_i, x_0) = 1$. According to [12], we can obtain a sparse kernel regression model based on the weight vector w .

Now we can extend RVR to multimodal RVR (M-RVR) for multimodal regression, by defining a new integrated kernel function for comparison of two multimodal data x and x_n as below:

$$k(x, x_n) = \sum_{m=1}^M c_m k^{(m)}(x^{(m)}, x_n^{(m)}) \quad (3)$$

Where $k^{(m)}$ denotes the kernel matrix over the m -th modality, similar to the definition given for the single modality case. This new integrated multiple-kernel can be expediently embedded into the conventional single-kernel RVR, and thus solved by the programs developed for the conventional single-kernel RVR. Here, we constrain the sum of c_m to be 1 and adopt a coarse-grid search through cross-validation on the training samples to find their optimal values.

2.2 Semi-supervised Multimodal RVR (SM-RVR)

Fig. 1 shows the flowchart of our proposed semi-supervised multimodal RVR (SM-RVR) method for multimodal regression. Here, the main idea is to select the most informative MCI subjects as unlabeled samples to aid the regression of AD and healthy controls (which are used as labeled samples). The algorithmic procedure of SM-RVR is detailed as below:

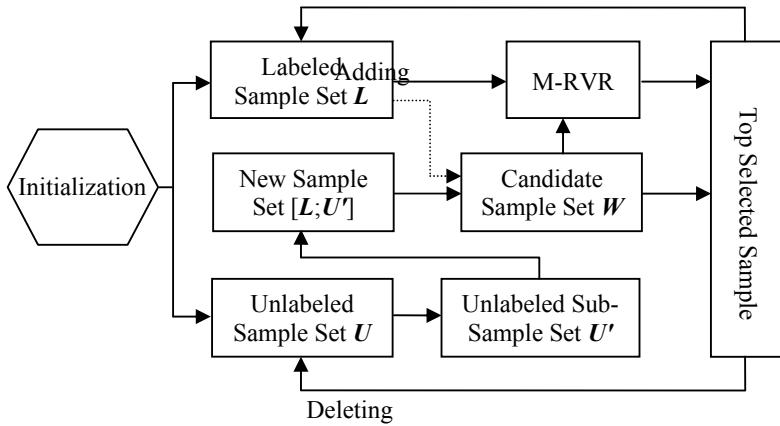
Step 1: Given a labeled sample set L and an unlabeled sample set U , initialize various parameters, including the maximum number of iterations T , RVR kernel function type, the kernel width parameter σ , and the number of the nearest neighbors k used in KNN algorithm (as detailed below);

Step 2: Randomly select n unlabeled samples from the unlabeled sample set U to constitute a new unlabeled sample set U' , and then combine U' with L to constitute a new sample set $[L; U']$ and further compute the respective graph Laplacian matrix I , where according to manifold learning theory [10,11], local smoothness is assumed between all (labeled and unlabeled) samples. From the graph Laplacian matrix I , select the top samples (from U') with the minimum distances to the labeled samples in L , as the candidate sample set W ;

Step 3: For each sample x_j in W , find its k -nearest neighbors in L , compute the mean of clinical values of those k neighbors as its estimated clinical score y_j , and train M-RVR on L plus $\{(x_j, y_j)\}$. Then, we compute the value of square root of mean square error (RMSE) for each sample x_j in W . Finally, a sample with the top confidence (i.e., minimum RMSE value) in W is selected and added into L , and further deleted from U .

Step 4: Go to Step 2 for running the next iteration;

Step 5: After reaching the total number of iterations T , train M-RVR on the latest L to build a final regression model.

**Fig. 1.** The flowchart of the proposed SM-RVR method

3 Results

3.1 Subjects

In this paper, the Alzheimer's disease Neuroimaging Initiative (ADNI) dataset is used to test our semi-supervised regression method. Only the baseline ADNI subjects with all corresponding MRI, PET, and CSF data are included, thus leading to a total of 202 subjects (including 51 AD patients, 99 MCI patients, and 52 healthy controls (HC)). Table 1 lists the demographics of these subjects.

The same image pre-processing as used in [3] is adopted here. First, for all structural MR images, we correct their intensity inhomogeneity by the N3 algorithm, do skull-stripping, and remove cerebellum. Then, we use the FSL package to segment each structural MR image into three different tissues: gray matter (GM), white matter (WM), and cerebrospinal fluid (CSF). We further use an atlas warping algorithm [16] to partition each structural MR brain image into 93 ROIs. For each of the 93 ROIs, we compute GM volume in that ROI as a feature; For PET image, we use a rigid transformation to align it to its respective structural MR image of the same subject, and then compute the average PET value of each ROI as a feature. Accordingly, for each subject, we can acquire 93 features from the structural MR image, another 93 features from the PET image, and 3 features from the CSF biomarkers.

Table 1. Subject information (mean \pm std)

	AD	MCI	HC
Age	75.2 \pm 7.4	75.3 \pm 7.0	75.3 \pm 5.2
Education	14.7 \pm 3.6	15.9 \pm 2.9	15.8 \pm 3.2
MMSE	23.8 \pm 1.9	27.1 \pm 1.7	29.0 \pm 1.2
ADAS-Cog	18.3 \pm 6.0	11.4 \pm 4.4	7.4 \pm 3.2

Table 2. Comparison of the regression performance of SM-RVR with respect to different combinations of MRI, PET and CSF modalities

Modality	MMSE		ADAS-Cog	
	RMSE	CORR	RMSE	CORR
MRI	2.1711	0.7306	5.1565	0.7004
PET	2.4613	0.6178	5.0406	0.7055
CSF	2.4493	0.6001	5.6168	0.6412
MRI+PET	2.0948	0.7551	4.7315	0.7621
MRI+CSF	2.0324	0.7709	4.9821	0.7376
PET+CSF	2.3828	0.6631	4.8907	0.7344
MRI+PET+CSF	1.9187	0.8013	4.4482	0.7823

3.2 Experimental Setup

To evaluate the performance of regression methods, we use both RMSE and correlation coefficient (CORR) [4] as performance measures. The number of the nearest neighbor k in KNN algorithm and the maximum number of iterations T are both learned from the training samples, through an enumeration search using the range from 1 to 50 and 1 to 99, respectively. For the performance evaluation of regression methods, we use a 10-fold cross-validation strategy to compute the average RMSE and CORR measures. The RVM regression learning machine is implemented using Sparse Bayesian toolbox¹, with Gauss kernel and default kernel-width σ . The weights in the M-RVR are learned based on the training samples, through a grid search using the range from 0 to 1 at a step size of 0.1. Also, for each modality feature f_i in labeled samples and unlabeled samples, the same feature normalization scheme as used in [3] is adopted here.

3.3 Experimental Results

Table 2 shows the performance measures (including RMSE and CORR) of our SM-RVR method, using different combinations of MRI, PET and CSF modalities. As we can see from Table 2, the combination of MRI, PET, and CSF can consistently achieve better results than any other methods. Specifically, SM-RVR using all three modalities can achieve a RMSE of 1.9187 and a CORR of 0.8013 for MMSE scores, and a RMSE of 4.4482 and a CORR of 0.7823 for ADAS-Cog scores, as shown in Fig. 2 which gives the scatter plots of actual clinical scores vs. estimated scores. On the other hand, Table 2 also indicates that the use of two modalities can improve the regression performance, although they are inferior to the use of all three modalities together. These results validate the advantage of multimodal regression over the conventional single-modal regression in estimation of clinical scores.

Table 3 shows the comparison of SM-RVR with supervised M-RVR. It is worth noting that, for fair comparison, we implement two versions of M-RVR, i.e., one using only AD and HC subjects as training sample and another using all (AD, HC and MCI) subjects as training samples. As can be seen from Table 3, SM-RVR consistently

¹ <http://www.miketipping.com/index.php?page=rvm>

outperforms M-RVR (including both versions) on each performance measure, which validates the efficacy of our SM-RVR method that uses MCI subjects only as unlabeled samples in a semi-supervised regression framework. Also, from Table 3, it is interesting to note that M-RVR using all subjects achieves slightly better performance in terms of RMSE, but much worse performance in terms of CORR, compared with M-RVR using only AD and HC subjects. This implies that the clinical scores of MCI subjects may contain more noises than those of AD or HC subjects.

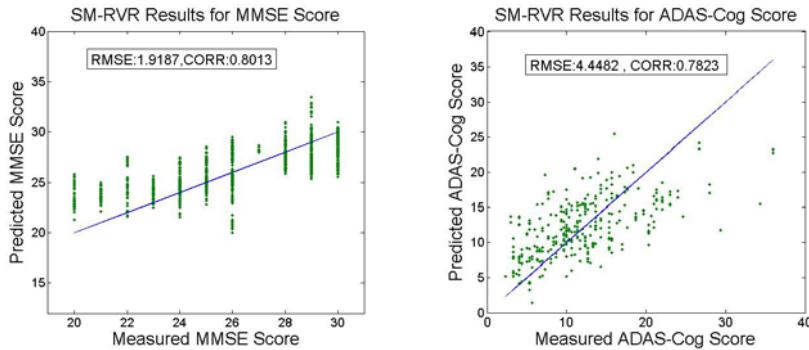


Fig. 2. Scatter plots of actual clinical scores vs. estimated scores for MMSE (left) and ADAS-Cog (right)

Table 3. Comparison of regression performance of SM-RVR and M-RVR

Methods	MMSE		ADAS-Cog	
	RMSE	CORR	RMSE	CORR
M-RVR (51AD+52HC)	2.2159	0.7285	4.9174	0.7325
M-RVR (51AD+52HC+99MCI)	2.1701	0.5261	4.6909	0.6404
SM-RVR(51AD+52HC+99MCI)	1.9187	0.8013	4.4482	0.7823

Finally, in Fig.3, we plot the curves of regression performance measures (of RMSE and CORR) with respect to the different number of unlabeled MCI samples when using different number of nearest neighbors (i.e., $k=1, 3, 5$) in SM-RVR. As can be seen from Fig. 3, the regression performance of SM-RVR is first steadily improved as the number of unlabeled MCI samples increases and is significantly better than that of M-RVR in most cases, but it declines after reaching a certain value. This implies that using selected MCI subjects as unlabeled samples is superior to using all MCI subjects as unlabeled samples in clinical score estimation.

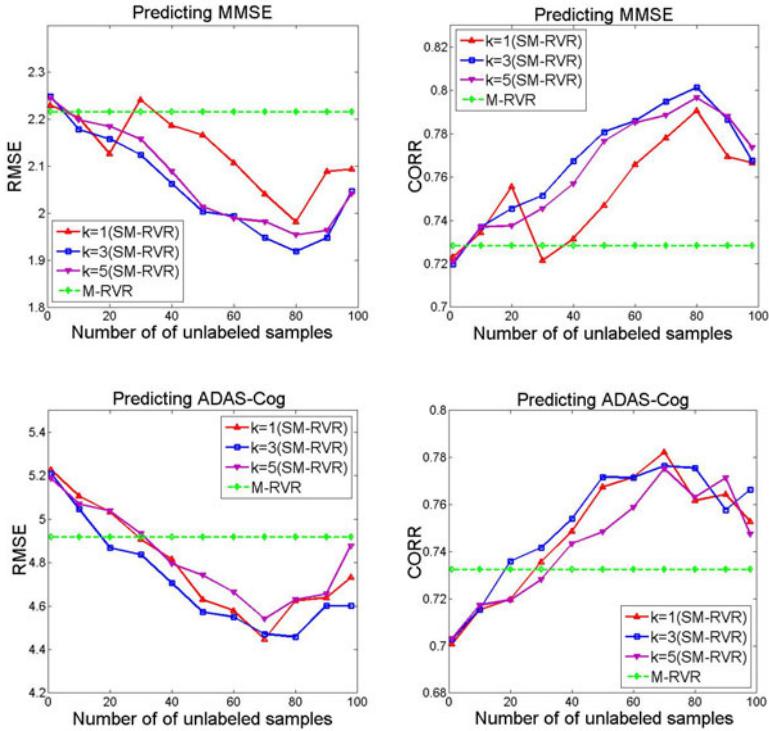


Fig. 3. Plots of regression performance (RMSE and CORR) vs. different number of unlabeled samples when using different number of nearest neighbors (i.e., $k=1, 3, 5$) in SM-RVR. The number of unlabeled MCI samples is equal to the total number of total iterations T .

4 Conclusion

This paper proposes a novel semi-supervised multimodal regression method, namely SM-RVR, to predict clinical scores of subjects (including AD, HC or MCI) from both imaging and biological biomarkers, i.e., MRI, PET and CSF. Our method assumes that the clinical scores obtained from MCI subjects may contain more noises than AD and HC subjects, and thus should be used in a semi-supervised regression framework as unlabeled data, rather than in a conventional supervised regression framework as labeled data. Furthermore, a scheme for selecting the most informative MCI subjects for helping training regression model is also derived. The experimental results on the ADNI dataset show the efficacy of our proposed method.

Acknowledgments. This work was supported in part by National Science Foundation of China under grant Nos. 60875030, 60973097 and 61035003, and also by NIH grants EB006733 and EB009634.

References

1. Stonnington, C.M., Chu, C., Klöppel, S., Jack, C.R., Ashburner, J., Frackowiak, R.S.J.: ADNI: Predicting clinical scores from magnetic resonance scans in Alzheimer's disease. *NeuroImage* 51(4), 1405–1413 (2010)
2. Wang, Y., Fan, Y., Bhatt, P., Davatzikos, C.: High-dimensional pattern regression using machine learning: From medical images to continuous clinical variables. *NeuroImage* 50(4), 1519–1535 (2010)
3. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D.: ADNI: Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage* 55(3), 856–867 (2011)
4. Fan, Y., Kaufer, D., Shen, D.: Joint estimation of multiple clinical variables of neurological diseases from imaging patterns. In: ISBI, pp. 852–855 (2010)
5. Walhovd, K.B., Fjell, A.M., Amlie, I., Grambaite, R., Stenset, V., Bjørnerud, A., Reinvang, I., Gjerstad, L., Cappelen, T., Due-Tønnesen, P., Fladby, T.: Multimodal imaging in mild cognitive impairment: Metabolism, morphometry and diffusion of the temporal-parietal memory network. *NeuroImage* 45(1), 215–223 (2009)
6. Ye, J., Chen, K., Wu, T., Li, J., Zhao, Z., Patel, R., Bae, M., Janardan, R., Liu, H., Alexander, G., Reiman, E.M.: Heterogeneous data fusion for Alzheimer's disease study. In: KDD 2008, pp. 1025–1033 (2008)
7. Fan, Y., Resnick, S.M., Wu, X., Davatzikos, C.: Structural and functional biomarkers of prodromal Alzheimer's disease: a high-dimensional pattern classification study. *NeuroImage* 41, 277–285 (2008)
8. Fjell, A.M., Walhovd, K.B., Fennema-Notestine, C., McEvoy, L.K., Hagler, D.J., Holland, D., Brewer, J.B., Dale, A.M.: CSF biomarkers in prediction of cerebral and clinical change in mild cognitive impairment and Alzheimer's disease. *J. Neurosci.* 30, 2088–2101 (2010)
9. Mosconi, L., Brys, M., Glodzik-Sobanska, L., De Santi, S., Rusinek, H., de Leon, M.J.: Early detection of Alzheimer's disease using neuroimaging. *Exp. Gerontol.* 42, 129–138 (2007)
10. Belkin, M., Niyogi, P., Sindhwani, V.: Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples. *Journal of Machine Learning Research* 7, 2399–2434 (2006)
11. Belkin, M., Niyogi, P.: Semi-Supervised Learning on Riemannian Manifolds. *Machine Learning* 56, 209–239 (2004)
12. Tipping, M.: Sparse Bayesian learning and the relevance vector machine. *J. Mach. Learn. Res.* 1, 211–244 (2001)
13. Filipovich, R., Davatzikos, C.: Semi-supervised pattern classification of medical images: Application to mild cognitive impairment (MCI). *NeuroImage* 55(3), 1109–1119 (2011)
14. Hinrichs, C., Singh, V., Xu, G., Johnson, S.: MKL for robust multi-modality AD classification. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5762, pp. 786–794. Springer, Heidelberg (2009)
15. Zhang, D., Shen, D.: Semi-supervised multimodal classification of Alzheimer's disease. In: ISBI, 1628–1631 (2011)
16. Shen, D., Davatzikos, C.: HAMMER: Hierarchical attribute matching mechanism for elastic registration. *IEEE Transactions on Medical Imaging* 21, 1421–1439 (2002)

Automated Cephalometric Landmark Localization Using Sparse Shape and Appearance Models

Johannes Keustermans¹, Dirk Smeets¹, Dirk Vandermeulen¹,
and Paul Suetens²

¹ Center for Processing Speech and Images, Department of Electrical Engineering,
Katholieke Universiteit Leuven, Belgium

² IBBT-K.U.Leuven Future Health department, Leuven, Belgium
johannes.keustermans@uz.kuleuven.ac.be

Abstract. In this paper an automated method is presented for the localization of cephalometric landmarks in craniofacial cone-beam computed tomography images. This method makes use of a statistical sparse appearance and shape model obtained from training data. The sparse appearance model captures local image intensity patterns around each landmark. The sparse shape model, on the other hand, is constructed by embedding the landmarks in a graph. The edges of this graph represent pairwise spatial dependencies between landmarks, hence leading to a sparse shape model. The edges connecting different landmarks are defined in an automated way based on the intrinsic topology present in the training data. A maximum a posteriori approach is employed to obtain an energy function. To minimize this energy function, the problem is discretized by considering a finite set of candidate locations for each landmark, leading to a labeling problem. Using a leave-one-out approach on the training data the overall accuracy of the method is assessed. The mean and median error values for all landmarks are equal to 2.41 mm and 1.49 mm, respectively, demonstrating a clear improvement over previously published methods.

1 Introduction

Cephalometric analysis studies the craniofacial morphology and enables both orthodontists and maxillofacial, craniofacial or plastic surgeons to make a diagnosis, as well as to plan and evaluate an appropriate treatment. It consists traditionally of identifying predefined anatomical landmarks on lateral cephalometric radiographs. Based on these anatomical landmarks relevant distance and angular measurements are performed. A major disadvantage of this technique is its inherent two-dimensional character. The recent introduction of Cone Beam Computed Tomography (CBCT) in daily practice enables a truly three-dimensional cephalometric analysis on a routine basis, due to its low radiation dose, low cost and possibility of scanning the patient in its natural head position [1,2]. Since manual localization of the cephalometric landmarks in CBCT images is a

very tedious approach, there is a strong need for automated methods. However, automated localization of these landmarks is hampered by the large variability in craniofacial morphology, certainly in pathological cases, and the presence of artifacts in the images.

In this paper we present an automated method for the localization of cephalometric landmarks in CBCT images. This method makes use of a statistical sparse appearance and shape model. For a review of existing 2D automated methods, see [3]. In [4], Keustermans *et al.* present an automated method for the localization of cephalometric landmarks in 3D CBCT images. However, a major shortcoming is the manual definition of the connections between the cephalometric landmarks. Furthermore, a combination of a local and global shape model is used leading to an iterative algorithm.

This paper is organized as follows. Section 2 describes the proposed algorithm. Section 3 presents the model training phase of the algorithm, whereas Section 4 handles the model fitting. Subsequently, Section 5 discusses the experimental results. Finally, Section 6 formulates a conclusion and some ideas for future work.

2 Method

We present a supervised automatic landmark localization algorithm that incorporates prior information on these cephalometric landmarks. The use statistical models are trained on CBCT images with manually annotated landmarks. The prior information is twofold: first, the image intensity patterns in the CBCT image around each individual landmark leading to a statistical model describing appearance, and second, the spatial relationships between the landmarks leading to a statistical model describing shape.

The statistical model describing appearance is built for each cephalometric landmark individually, considering only a local region around the landmark in the CBCT image, hence leading to a sparse appearance model. Local image descriptors are used to describe the image intensity patterns in this local region. To build the statistical model describing shape, the landmarks are connected by edges, leading to a graph representation where the edges represent pairwise dependencies between landmarks. The edges are determined in an automated way based on the intrinsic graph topology present in the training data. Therefore, we search for dependencies among the cephalometric landmark positions by computing the description lengths of multivariate Gaussian models. By imposing the markovianity property onto the graph, each landmark is directly dependent upon its neighboring landmarks only, hence leading to a sparse shape model. The advantage of sparse shape models is the increased flexibility compared to global shape models. On the other hand, only pairwise dependencies can be modeled.

2.1 Bayesian Inference

The goal of the algorithm is the optimal localization of a number of predefined cephalometric landmarks. The accuracy of the solution can be expressed as a conditional probability on the location of the landmarks \mathcal{L} given the image \mathcal{I} .

Using bayesian inference, this conditional probability can be reformulated as $P(\mathcal{L}|\mathcal{I}) = \frac{P(\mathcal{I}|\mathcal{L})P(\mathcal{L})}{P(\mathcal{I})}$. Here, $P(\mathcal{I}|\mathcal{L})$ is the image likelihood and $P(\mathcal{L})$ is the shape prior giving rise to, the sparse appearance model and the sparse shape model. For the remaining of this paper probabilities are converted into energies by taking the negative logarithm, leading to the following energy function to be minimized

$$\mathcal{L}^* = \arg \min_{\mathcal{L}} (E_I(\mathcal{I}, \mathcal{L}) + E_S(\mathcal{L})) , \quad (1)$$

where $E_I(\mathcal{I}, \mathcal{L})$ and $E_S(\mathcal{L})$ are the negative logarithm of $P(\mathcal{I}|\mathcal{L})$ and $P(\mathcal{L})$, respectively.

2.2 Sparse Appearance Model

The image likelihood is simplified by making two assumptions. First, we assume that the influence of the presence of a landmark on the image is only local, justifying the use of local image descriptors to capture the local intensity patterns around each landmark. Second, we assume that local image descriptors are mutually independent for each landmark. In this paper, the nSIFT image descriptor is used [5]. This descriptor consists of a set of gradient histograms over a window centered on each landmark.

Using the aforementioned assumptions the image likelihood term can be rewritten as follows

$$P(\mathcal{I}|\mathcal{L}) = \prod_i^n P(\mathcal{I}|l_i) = \prod_i^n P_i(\omega_i) . \quad (2)$$

In this equation l_i represents the i^{th} anatomical landmark, n is the number of landmarks and ω_i represents the local image descriptor corresponding to landmark l_i . Finally, converting the probabilities into energies by taking the negative logarithm yields the following expression:

$$E_I(\mathcal{I}, \mathcal{L}) = \sum_i^n d_i(\omega_i) , \quad (3)$$

where d_i is the energy function corresponding to landmark i .

2.3 Sparse Shape Model

The shape prior introduces the sparse shape model into the Bayesian framework. Two assumptions are made as well. First, invariance of the shape model with respect to translations is assumed. Second, we assume that a landmark directly interacts with its neighbors only, thereby imposing the sparse nature of the shape model. The sparse shape model is constructed by embedding the landmarks in an undirected graph \mathcal{G} . The undirected graph $\mathcal{G} = \{\mathcal{V}, \mathcal{E}\}$ consists of a set of nodes $\mathcal{V} = \{l_1, \dots, l_n\}$, the landmarks, and a set of edges \mathcal{E} . Let \mathcal{N} be a neighborhood

system defined on \mathcal{V} , where $\mathcal{N}_i = \{l_j \in \mathcal{V} | (l_i, l_j) \in \mathcal{E}\}$ denotes the set of neighbors of l_i . A clique c of the graph \mathcal{G} is a fully connected subset of \mathcal{V} . The set of cliques is represented by \mathcal{C} . For each $l_i \in \mathcal{V}$, let X_i be a random variable taking values x_i in some discrete or continuous sample space \mathcal{X} . By concatenating the variables at each node, we obtain a random vector $\mathbf{X} = \{X_i | l_i \in \mathcal{V}\}$. By the Hammersley-Clifford theorem a Markov Random Field can be defined in terms of a decomposition of the distribution over cliques of the graph:

$$p(\mathbf{x}) = \frac{1}{Z} \exp \left(- \sum_{c \in \mathcal{C}} V_c(\mathbf{x}) \right), \quad (4)$$

where Z represents the partition function and V_c represents the clique potential function. In the algorithm presented in this paper only pairwise cliques are used. Converting probabilities into energies and taking into account the translation invariance leads to the following energy function for the local shape model

$$E_S(\mathcal{L}) = \sum_{(i,j) \in \mathcal{E}} d_{ij}(\mathbf{x}_j - \mathbf{x}_i), \quad (5)$$

where d_{ij} is the energy function of the edge between the landmarks l_i and l_j and corresponds to the clique potential function.

2.4 Edge Definition

The pairwise connections or edges between the cephalometric landmarks encode spatial dependencies in the sparse shape model. To make the sparse shape model as robust as possible, only sensible edges should be defined. This is, however, a complex task, highly dependent upon the training data. To analyze the intrinsic spatial dependencies between the cephalometric landmarks in the training data, we compute the description length of a multivariate Gaussian model for each pairwise combination of landmarks. The description length quantifies the compactness of the model capturing the edge variation. The rationale behind the compactness is the assumption that a high spatial dependency between two landmarks leads to a more compact model. The description length comprises the cost (number of bits) of communicating a model itself and the data encoded with the model. For more details we refer to Davies *et al.* [7]. In this framework we use a simplified version of the description length as presented by Thodberg [8]. This approach is similar to the method presented by Langs *et al.* [9]. In this paper, however, only pairwise subsets are used and no alignment of the subsets is performed. The use of a multivariate Gaussian edge model constrains the pairwise dependencies to be linear. However, this approach can be extended to nonlinear dependencies as well, using kernel principal component analysis [6] (see Section 3).

Figure 1 gives the result of applying this method to the training data of the cephalometric landmarks. From this figure it can be seen that the left-right

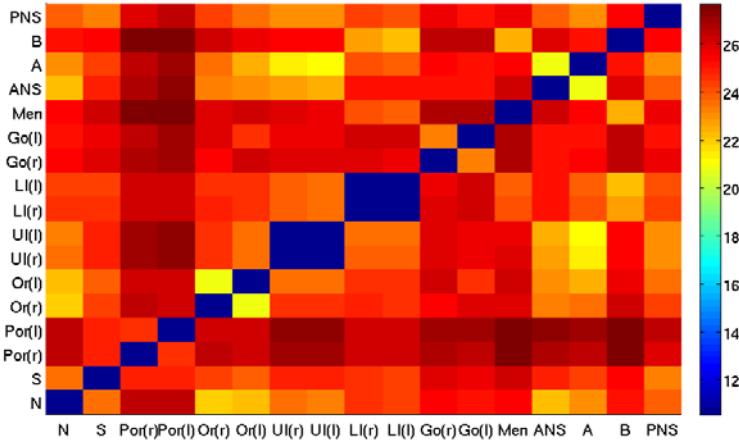


Fig. 1. Description length for all pairwise connections between the cephalometric landmarks. Blue indicates a description length of value 10 or lower, whereas red indicates a description length of value 26 or higher.

connections between the landmarks have a low description length (*upper incisor* (UI), *lower incisor* (LI), *orbitale* (Or) and *gonion* (Go)). Furthermore, the description length of any connection to the *porion* (Por) is high. This is caused by artifacts present in the CBCT at the level of this landmark, hampering the correct localization of this landmark.

Given the description length for all pairwise connections, edges need to be defined. Therefore we apply a weight to each edge, depending on its description length. An edge with a low description length will be assigned a high weight, whereas an edge with a high description length will be assigned a low weight, making it less important in the sparse shape model. For the assignment of the weights, the sigmoid function is used.

3 Model Training

For the sparse appearance and shape model, probability density functions need to be estimated from the training data. For the sparse shape model a Gaussian distribution provides good results. For the sparse appearance model, on the other hand, a more advanced method is needed, which is related to kernel principal component analysis [6]. It consists of a nonlinear mapping $\phi : \mathbb{R}^d \mapsto F$ of the training samples, which are the nSIFT feature descriptors in each landmark, to a higher (possibly infinite) dimensional feature space F in which a Gaussian distribution is presumed. There is no need to explicitly compute the nonlinear mapping ϕ , since only scalar products in feature space are needed that, by use of the Mercer theorem, can be evaluated using a positive definite kernel function $k(\omega_i, \omega_j) = \langle \phi(\omega_i), \phi(\omega_j) \rangle$. The Gaussian kernel is used with a bandwidth equal to the mean nearest-neighbor training sample distance.

4 Model Fitting

The actual localization of the landmarks in a CBCT image comes down to optimizing the energy function (1). This is not a straightforward operation since the energy function is not convex. However, due to the recent advances in solving the inference problem in multilabel Markov Random Fields the non-convexity can be partially solved by a discretization of the energy function.

4.1 Discretization

The discretization of the energy function is performed by imposing a discrete sample space \mathcal{X} on the graph \mathcal{G} . This discrete sample space consists of a finite set of candidate landmark locations. As such, the optimization problem comes down to the selection of the most optimal candidate location for each landmark. These candidate locations are obtained by evaluating the image likelihood $d_i(\omega_i)$ in a search grid located around the expected location of the landmark of interest and selecting the m locations with lowest cost. This results in a set of candidates $\mathbf{x}_i = \{\mathbf{x}_{ik}\}_{k=1}^m$ for each landmark. Important to note is that we assume that the CBCT image is rigidly registered to the images in the training data set. The optimization problem now becomes a labeling problem: $\mathbf{r} = \{\mathbf{r}_i \in \{0, 1\}^m\}$. Following conditions must hold: $\sum_{k=1}^m r_{ik} = 1$ and $r_{ik} = 1$ if candidate k is selected. The resulting discrete cost function to be minimized becomes

$$\arg \min_{\mathbf{r}} \left(\sum_{i=1}^n \sum_{k=1}^m r_{ik} d_i(\omega_{ik}) + \gamma \sum_{i=1}^n \sum_{k=1}^m \left(r_{ik} \sum_{j=1}^n \sum_{o=1}^m r_{jo} d_{ij}(\mathbf{x}_{ik}, \mathbf{x}_{jo}) \right) \right), \quad (6)$$

where γ is a constant that determines the relative weight of the image likelihood and the shape prior.

4.2 Markov Random Field Optimization

Currently two types of methods are most prominent in multilabel Markov random field optimization: the methods based on graph cuts [10] and those based on message-passing [11,12]. However, the methods based on graph cuts can not be used to minimize the energy function (6) because it is not submodular [13]. Here a variant of the Belief Propagation (BP) algorithm is used. Belief propagation is a local message passing algorithm that converges to a fixed point for an acyclic graph. Since the graphs obtained by the method in this paper in general do contain cycles, we use the BP-fusion algorithm of Lempitsky *et al.* [14]. This BP-fusion algorithm combines different belief propagation proposals using the fusion move algorithm.

5 Experiments and Results

The algorithm described in this paper is used to identify 17 cephalometric landmarks in CBCT images. These landmarks are *nasion*, *sella*, *porion* (left and

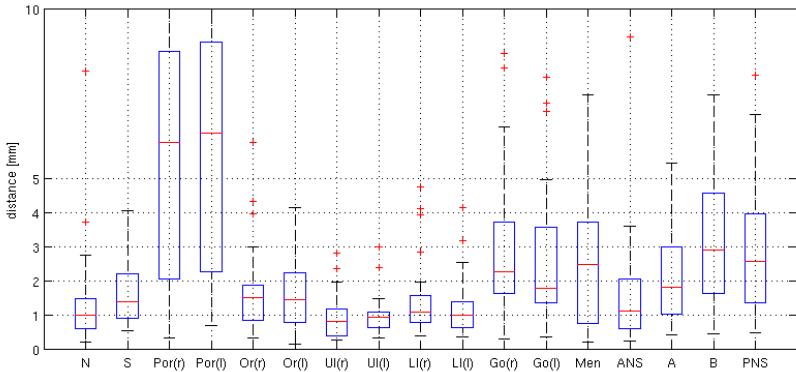


Fig. 2. Box plot of the errors for each landmark using a leave-one-out approach on the training data

right), orbitale (left and right), upper incisor (left and right), lower incisor (left and right), gonion (left and right), menton, anterior nasal spine, A-point, B-point and posterior nasal spine [1]. The training data set consisted of 37 CBCT images ($0.4 \times 0.4 \times 0.4$ mm) from patients undergoing maxillofacial surgery, with manually identified landmarks. A leave-one-out approach is used to validate the algorithm.

Figure 2 shows the results of the leave-one-out approach. The mean and median error values for all landmarks are equal to 2.41 mm and 1.49 mm respectively. Compared to the results obtained by Keustermans *et al.* [4], there is an improvement of 0.14 mm and 0.23 mm for the mean and median error respectively. The median error values separated in the transversal, sagittal and transversal direction are equal to 0.58 mm, 0.55 mm and 0.61 mm, respectively. This allows to compare the results to the inter- and intra-observer variability of manual landmark localization, which is equal to respectively 0.78 mm, 0.86 mm and 1.26 mm for the inter-observer variability and 0.88 mm, 0.76 mm and 0.84 mm for the intra-observer variability, as reported in [1].

6 Conclusion

A statistical model-based algorithm for the localization of cephalometric landmarks in CBCT images incorporating both a sparse appearance and shape model is presented. The sparse appearance model is constructed through the use of local image descriptors. The sparse shape model is obtained by embedding the landmarks in a graph. The edges of the graph are defined based on the intrinsic dependencies between the landmark positions present in the training data. The energy function, obtained from a maximum a posteriori approach, is optimized efficiently using multilabel Markov Random Fields. The accuracy of this method is shown using a leave-one-out approach on the training data, demonstrating an

outperformance over existing methods. Since the training data are obtained from patients undergoing maxillofacial surgery, the robustness of the algorithm with respect to pathologies is shown.

Acknowledgments. The authors gratefully acknowledge the financial support by K.U.Leuven Concerted Research Action GOA/11/006 and Nobel Biocare.

References

1. Swennen, G.R.J., Schutyser, F., Hausamen, J.-E.: Three-Dimensional Cephalometry, A Color Atlas and Manual. Springer, Heidelberg (2006)
2. Swennen, G.R.J., Schutyser, F., Barth, E.L., De Goeve, P., De Mey, A.: A New Method of 3-D Cephalometry Part I: The Anatomic Cartesian 3-D Reference System. *Journal of Craniofacial Surgery* 17(2), 314–325 (2006)
3. Leonardi, R., Giordano, D., Maiorana, F., Spampinato, C.: Automatic Cephalometric Analysis. *Angle Orthod.* 78(1), 145–151 (2009)
4. Keustermans, J., Mollemans, W., Vandermeulen, D., Suetens, P.: Automated Cephalometric Landmark Identification using a Local Shape and Appearance Model. In: Proc. ICPR (2010)
5. Cheung, W., Hamarneh, G.: n-SIFT: n-dimensional scale invariant feature transform. *IEEE Trans. Image Process.* 18(9), 2012–2021 (2009)
6. Schölkopf, B., Smola, A., Müller, K.R.: Nonlinear component analysis as a kernel eigenvalue problem. *Neural Computation* 10, 1299–1319 (1998)
7. Davies, R.H., Twining, C.J., Cootes, T.F., Waterton, J.C., Taylor, C.J.: A minimum description length approach to statistical shape modelling. *IEEE Trans. Med. Imaging* 21(5), 525–537 (2002)
8. Thodberg, H.H.: Minimum description length shape and appearance models. In: Taylor, C.J., Noble, J.A. (eds.) IPMI 2003. LNCS, vol. 2732, pp. 51–62. Springer, Heidelberg (2003)
9. Langs, G., Paragios, N.: Modeling the structure of multivariate manifolds: Shape maps. In: Proc. CVPR (2008)
10. Boykov, Y., Veksler, O., Zabih, R.: Fast Approximate Energy Minimization via Graph Cuts. *IEEE Trans. Pattern Anal. Mach. Intell.* 23(11), 1222–1239 (2001)
11. Wainwright, M.J., Jaakkola, T.S., Willsky, A.S.: MAP Estimation Via Agreement on Trees: Message-Passing and Linear Programming. *IEEE Trans. Inf. Theory* 51(11), 3697–3717 (2006)
12. Kolmogorov, V.: Convergent Tree-reweighted Message Passing for Energy Minimization. *IEEE Trans. Pattern Anal. Mach. Intell.* 28(10), 1568–1583 (2006)
13. Kolmogorov, V., Zabih, R.: What Energy Functions Can Be Minimized via Graph Cuts? *IEEE Trans. Pattern Anal. Mach. Intell.* 26(2), 147–159 (2001)
14. Lempitsky, V., Rother, C., Roth, S., Blake, A.: Fusion Moves for Markov Random Field Optimization. *IEEE Trans. Pattern Anal. Mach. Intell.* 32(8), 1392–1405 (2010)

A Comparison Study of Inferences on Graphical Model for Registering Surface Model to 3D Image

Yoshihide Sawada and Hidekata Hontani

Nagoya Institute of Technology

Abstract. In this article, we report on a performance comparison study of inferences on graphical models for model-to-image registration. Both Markov chain Monte Carlo (MCMC) and nonparametric belief propagation (NBP) are widely used for inferring marginal posterior distributions of random variables on graphical models. It is known that the accuracy of the inferred distributions changes according to the methods used for the inference and to the structures of graphical models. In this article, we focus on a model-to-image registration method, which registers a surface model to given 3D images based on the inference on a graphical model. We applied MCMC and NBP for the inference and compared the accuracy of the registration on different structures of graphical models. Then, MCMC outperformed NBP significantly in the accuracy.

Keywords: registration, Markov chain Monte Carlo, nonparametric belief propagation, graphical model.

1 Introduction

Markov random fields (MRF) [1] or undirected graphical models [1] are widely used for representing statistics of targets and have variety of applications [2,3]. When a graphical model is used, given a set of measurements, we can obtain a solution for the problem by inferring the probability distributions of the random variables on the graphical model. In many cases, it is not easy to infer the exact distributions, and it is feasible to use some effective approximation method. This is why many inference methods, such as Markov chain Monte Carlo (MCMC) [1] or belief propagation (BP) [1,3], are broadly employed.

In this article, we focus on a model-to-image registration method proposed by Hontani and Watanabe [2]. In the method, a point distribution model (PDM) is employed for representing a surface of a target organ and the statistics of the points is represented by a graphical model. The method registers the model to given 3D images by inferring the marginal posterior distribution of the coordinates of each of the points. The advantage of this method is that the confidence of the registration can be evaluated at each point on the surface. Such evaluation of confidence is essential not only for registration but also for other autonomous applications [4]. The method uses the non-parametric belief propagation (NBP) [2,3] for the inference.

However, it is known that the estimates inferred by a BP can be biased when the graphical model is cyclic and that the accuracy of the estimates would change depending on the structure of the graph [5]. This means that the performance can change depending

on the inference method, even when the graphical model and the given measurements are identical. In addition, the method determines the structure of the graphical model of the target surface without any statistical analysis of the random variables. It is not only the method [2] but also other methods [3,6] determine the structure of graphical model without the statistical analysis.

In this article, we report on a performance comparison study of MCMC and NBP, which are widely used for the inference, in the model-to-image registration. We found that MCMC is much better than NBP.

2 Surface Registration with Graphical Model

In this article, we focus on a non-rigid surface registration method proposed in [2]. The method registers a surface model of a target organ to given X-CT images. The surface model is represented by a set of points, and is registered to given images by inferring the marginal posterior distribution of the location of each of the points. In the following, some details of the registration method is described.

2.1 Surface Model Construction

A surface is represented with a set of N points $\{P_j\}$ ($j = 1, 2, \dots, N$). Let \mathbf{x}_j denote a 3D coordinates of P_j . A statistical model of the points is constructed based on a sets of M X-CT images. Let $\mathcal{I} = \{I^i | i = 1, 2, \dots, M\}$ denote the set of training images. These images are normalized based on the body shape in advance. Let S^i denote the surfaces of the target organ extracted from I^i , manually.

For constructing the probabilistic model, M sets of N corresponding points $\{P_j^i\}$ are distributed on $\{S^i\}$. For obtaining this corresponding points, the entropy-based particle system [7] is used in the method [2]. Based on the resultant point sets $\{P_j^i\}$, we compute three probability distributions that obey the normal distributions [2]: The prior probability of the location of P_j , $p(\mathbf{x}_j)$, the probability of the local appearance I_j around P_j , $p(I_j|\mathbf{x}_j)$, and the probability distribution of the relative locations between two neighboring points P_j and P_k , $p(\mathbf{x}_j - \mathbf{x}_k)$. Using these probabilities, the simultaneous probability distribution is represented as follows:

$$p(\{\mathbf{x}_j\}, \{I_j\}) = \frac{1}{Z} \prod_j \psi_j(\mathbf{x}_j, I_j) \prod_{e_{j,k} \in \mathcal{E}} \psi_{j,k}(\mathbf{x}_j, \mathbf{x}_k), \quad (1)$$

where $\psi_j(\mathbf{x}_j, I_j) = p(\mathbf{x}_j)p(I_j|\mathbf{x}_j)$ and $\psi_{j,k}(\mathbf{x}_j, \mathbf{x}_k) = p(\mathbf{x}_j - \mathbf{x}_k)$.

For which $p(\mathbf{x}_j - \mathbf{x}_k)$ are represented, we have to determine a set of pairs (j, k) . In other words, we have to generate a set of edges $\mathcal{E} = \{e_{j,k}\}$. In [2], two nodes \mathbf{x}_j and \mathbf{x}_k are linked with an edge when the average distance between P_j and P_k ($i = 1, 2, \dots, M$) is shorter than a threshold. Strictly speaking, two nodes in a graphical model should be linked if and only if the corresponding two variables are not conditionally independent. Though, it is not trivial to learn appropriate structures of graphical models, especially when the number of nodes is much larger than the number of training data. There are some methods for estimating the structures of graphical models [8,9].

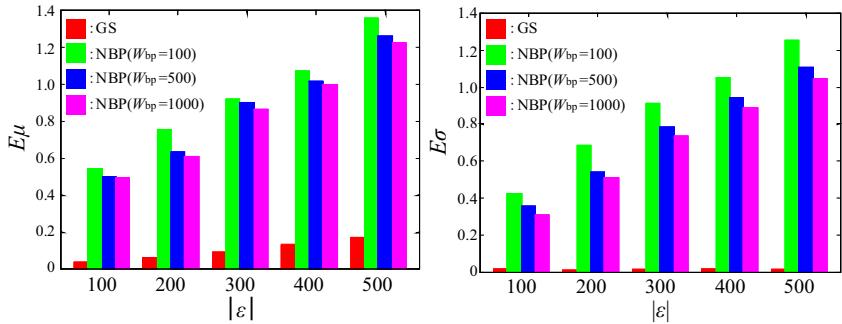


Fig. 1. Quantitative comparison of errors in inferences. Left: error in $\hat{\mu}_i$. Right: error in $\hat{\sigma}_i$.

Many of them estimate the structures by estimating the dependences among the variables with some regularization, which makes the structure more sparse for improving the generalization error. In this article, we determine the structure in a direct manner: We compute the canonical correlation $c_{j,k}$ [10] between any two variables x_j and x_k , and connect them with an edge if the correlation is enough large. This is just an approximation method for estimating the structures, but the advantage of this method is that we can easily control the number of edges for the comparison study.

2.2 Registration of Surface Model

Given a new X-CT image, we register the surface model to it by inferring the marginal posterior distribution $p_m(\mathbf{x}_j)$ of each point of the model. For this inference, we can use both the Gibbs sampling (GS) [1], and the non-parametric belief propagation (NBP) [2,3]. GS is a widely applicable MCMC algorithm, and in this article, we keep GS until a set of all samples satisfies the condition in [11]. The registration algorithm is as follows [2]:

1. Given a new X-ray CT image, we register the point distribution model by inferring the marginal posterior distributions $p_m(\mathbf{x}_j)$ ($j = 1, 2, \dots, N$) of each of the points.
 - (a) Normalize a body shape in the image.
 - (b) Compute $p(\mathbf{x}_j)p(I_j|\mathbf{x}_j)$ as the initial state.
 - (c) Apply GS or NBP for inferring the marginal posterior distribution $p_m(\mathbf{x}_j)$.
2. We estimate the location of the surface by computing the averages of $p_m(\mathbf{x}_j)$ and the confidence of the estimates by computing the variance of $p_m(\mathbf{x}_j)$.

It should be noted that the distribution $p(\mathbf{x}_j)p(I_j|\mathbf{x}_j)$ does not obey the Gaussian because the distribution $p(I_j|\mathbf{x}_j)$ is not the Gaussian [2]. This is why we cannot apply the parametric methods, such as BP.

3 Simulation Study of GS and NBP

For comparison purpose, we generated a set of various graphical models, on which we can compute the analytic solution of the marginal posterior distribution on each variable. We applied GS and NBP for inferring the distributions on each graphical model, and compared the resultant distributions with the analytic ones.

Let x_i ($i = 1, 2, \dots, N$) denote a one-dimensional random variable, and let $e_{i,j}$ denote the edge that connects x_i and x_j in the graph. In this study, we set every probability distribution as the Gaussian: $p(x_i) = \mathcal{N}(x_i | \mu_i, \alpha)$ and $p(x_i | x_j) = p(x_j | x_i) = p(x_i - x_j) = \mathcal{N}(x_i - x_j | 0, \beta)$, where μ_i is the mean of x_i , α is the variance of x_i , and β is that of $x_i - x_j$. It should be noted that the values of α and β are constant to all nodes and edges, respectively. Then, the marginal posterior distribution $p_m(x_i)$ obeys the Gaussian because the simultaneous probability distribution

$$p(\{x_i\}) = \frac{1}{Z} \prod_i \exp\left(-\frac{\alpha}{2}(x_i - \mu_i)^2\right) \prod_{e_{i,j} \in \mathcal{E}} \exp\left(-\frac{\beta}{2}(x_i - x_j)^2\right), \quad (2)$$

is also Gaussian. Let the mean and the variance of the marginal posterior distribution $p_m(x_i)$ be denoted as $\bar{\mu}_i$ and $\bar{\sigma}_i^2$. Then we can compute $\bar{\mu}_i$ and $\bar{\sigma}_i$ easily [12].

In the simulation experiments, fixing the number of the nodes as $N = 50$ and setting the number of edges $|\mathcal{E}|$, we randomly connected the nodes with edges to generate variety of graph structures. We generated a graph for each value of $|\mathcal{E}| = 100, 200, 300, 400$, and 500 . Determined the mean value μ_i randomly, we computed the analytic solution of $\bar{\mu}_i$ and $\bar{\sigma}_i$ [12].

Then, we inferred the marginal posterior distributions by NBP and GS. Figure 1 shows the errors of the estimates. The top row in the figure shows $E_\mu = \sum_i |\hat{\mu}_i - \bar{\mu}_i|/N$ and the bottom row shows $E_\sigma = \sum_i |\hat{\sigma}_i - \bar{\sigma}_i|/N$, where $\hat{\mu}_i$ and $\hat{\sigma}_i$ denote the estimates, and W_{bp} is the number of the Gaussian Components for NBP [2,3]. As shown in this figure, we can see that GS inferred more accurately than NBP, and that the distributions estimated by the NBP became more accurate when the number of the Gaussians, W_{bp} , was increased.

4 Comparison of the Registration Accuracy

In this section, we experimentally compared GS and NBP in terms of the registration accuracy. In addition, the structures of the graphical models were also compared in terms of the registration accuracy. We selected the aorta and the liver for target organs.

4.1 Construction of Statistical Model

We used $M = 15$ X-CT training images and $M_t = 15$ test images for the aorta, and $M = 26$ and $M_t = 8$ ones for the liver. From each image I^i , an expert manually traced the surface of the target organ. The traced surface was used for a training surface S^i . We applied the entropy-based particle system [7], and obtained M sets of corresponding points $\{P_j^i\}$. We generated $N = 200$ corresponding points on each surface of the aorta, and $N = 500$ points on the liver. Figure 2(a) and (b) show examples of the corresponding points distributed on the aorta arches.

As we mentioned in sec. 2.1, we computed the canonical correlations $c_{j,k}$ for all pairs of the points P_j and P_k . The distribution of $c_{j=j_0,k}$ ($k = 1, 2, \dots, N, k \neq j_0$) is shown in Fig. 2(c), in which P_{j_0} is pointed by an black arrow. As shown in the figure, the points neighboring to P_{j_0} have stronger correlations with P_{j_0} . In addition, we found that some points also had strong correlations with some distant points indicated by the white arrow.

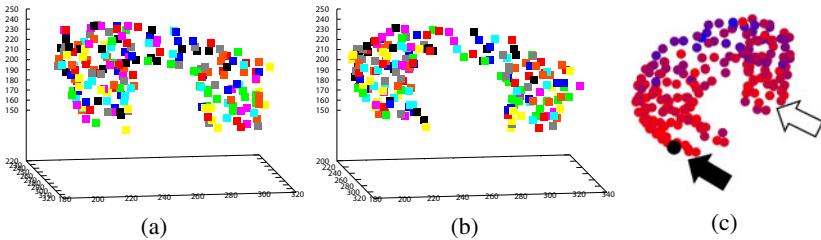
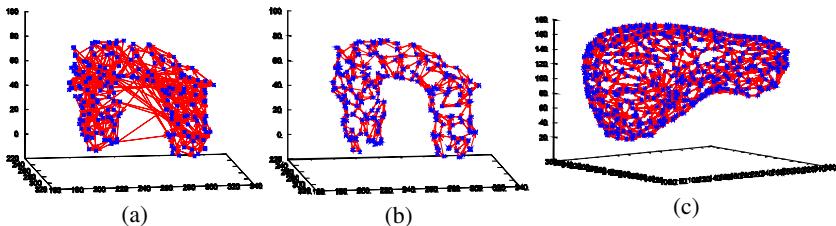


Fig. 2. (a),(b):Examples of corresponding points. The color represents the number of point. (c):Distribution of canonical correlations. The point P_{j_0} is indicated by a black arrow. The white arrow indicates a region that is highly correlated with P_{j_0} .



For comparison purpose, we determined the structure of the graphical model in two ways: (1) linking nodes based on their canonical correlations, and (2) linking nearest neighboring nodes as the method in [2]. After we determined the structure, the potential of the relative positions $\psi_{i,j} = p(\mathbf{x}_i - \mathbf{x}_j)$ is estimated for each edge $e_{i,j}$.

We inferred the marginal posterior distributions $p_m(\mathbf{x}_j)$ by means of GS and NBP. Let the expectation of the location \mathbf{x}_j^i be denoted by $\hat{\mu}_j^i$, which is computed from $p_m(\mathbf{x}_j)$. We evaluated the accuracy of the registration based on the residuals $E^i = \sum_j |\hat{\mu}_j^i - \bar{\mu}_j^i|/N$, where $\bar{\mu}_j^i$ denotes a point on S^i that is the closest to $\hat{\mu}_j^i$.

4.2 Comparison Study of GS and NBP

Firstly, determining the structure of the graphical model based on the canonical correlations, we compared the registration accuracy of the aorta between the GS and the NBP. By setting the degree of the graph $n = 3$, we constructed a graphical model shown in Fig. 3(a). As shown in Fig. 3(a), some pairs of two distant points were connected by edges in the graphs of the aorta. This means that the locations of these points are highly correlated each other even though they are distantly located each other. It should be noted that those pairs were not linked when only the neighboring point pairs were linked as in the method of [2] (Fig. 3(b)).

Figure 4 shows an example of the distributions of $p(\mathbf{x}_j)p(I_j|\mathbf{x}_j)$ and marginal posterior distributions $p_m(\mathbf{x}_j)$ inferred by GS and NBP ($W_{bp} = 1000$). As shown in this figure, the variance of each of the points decreased. This means that the confidence

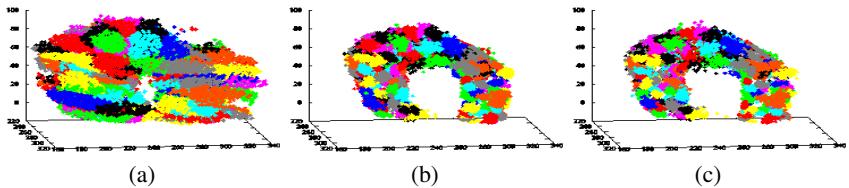


Fig. 4. An example of changes of distributions. (a): The initial distributions $p(\mathbf{x}_i)p(I_i|\mathbf{x}_i)$ (b), (c): The marginal posterior distributions $p_m(\mathbf{x}_j)$ inferred by GS and NBP.

Table 1. Errors of registered aorta surfaces

	min(pixel)	average(pixel)	max(pixel)
GS	1.60	2.71	4.94
NBP	2.37	3.86	6.34

Table 2. Errors of registered liver surfaces

	min(pixel)	average(pixel)	max(pixel)
GS	3.61	5.70	13.9
NBP	4.84	6.34	13.8

of the estimates increased by the inference on the graphical model by GS and NBP. It should be noted that the models of the relative positions $p(\mathbf{x}_j - \mathbf{x}_k)$ are not used before the inference on the graphical model. We found that both NBP and the GS increased the confidence of the estimated distributions.

Table 1 shows the minimum, the average, and the maximum values of E^i of each test image. As shown in the result, GS registered the model more accurately than NBP. We applied the Wilcoxon signed rank-sum test and the null hypothesis was rejected ($p < 0.01$): In this case, we can surely decide that GS outperformed NBP.

Next, we compared the registration accuracy of the liver between the GS and the NBP ($W_{bp} = 2000$). By setting the degree of the graph $n = 3$, we constructed a graphical model shown in Fig. 3(c). We inferred the marginal posterior distributions $p_m(\mathbf{x}_j)$ on the graph by using GS and NBP. Table 2 shows results. GS registered the liver model more accurately than NBP, as in the case of the aorta. The Wilcoxon signed rank-sum test ($p < 0.05$) rejected the null hypothesis and the superiority of the GS was accepted.

We also compared the run-time efficiency. In the case of registering aortas, NBP was about three times faster than GS.

4.3 Comparison Study of Structures of Graphical Models

We compared the registration accuracy of the aorta in terms of the structures of the graphical models. Changing the degree of the graph n , we obtained five different structures of the graphs.

For the comparison purpose, we inferred $p_m(\mathbf{x}_j)$ on all graphical models by means of GS and NBP, and estimated the deviation E^i . The graph of the average of E^i is shown in Fig. 5. The black bars indicate the minimum and the maximum of E^i .

As shown in the graph, GS outperformed NBP at each n . In addition, we can decide that the followings are right based on the Wilcoxon signed rank-sum tests ($p < 0.01$): When the degree of the graph n increased, the performance of NBP became worse but that of GS did not become worse.

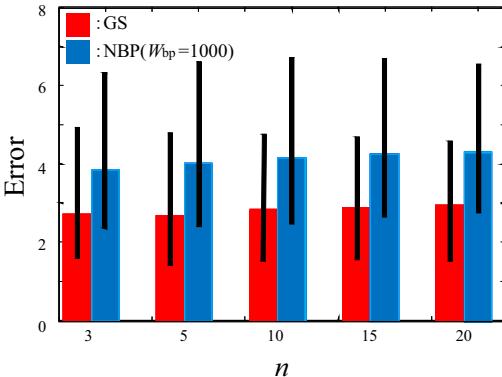


Fig. 5. The registration errors E^i and the degree n of the graph. Red: GS. Blue: NBP

Next, we compared the graphical model shown in Fig. 3(a) with other graphical models. For this comparison, setting the degree of the graph $n = 3$, we determined another structure of the graphical model by linking each point with $n = 3$ nearest neighbor points as in the method [2]. Figure 3(b) shows the resultant graph, which has no edge that connects distant two points.

Using GS, we inferred the marginal posterior distribution $p_m(x_j)$ on the two graphs that have the same degree of the graph $n = 3$: One graphical model was made by the canonical correlation-based method and the other was made by the nearest neighbor-based method. Let the former graph be denoted by \mathcal{G}_{CC} and the latter one be denoted \mathcal{G}_{NN} .

Table 3 shows the registration errors. It looks that the inference on \mathcal{G}_{CC} is more accurate than that on \mathcal{G}_{NN} . We did the Wilcoxon signed rank-sum test ($p < 0.01$) and it rejected the null hypothesis: We can decide that the accuracy of the inference on \mathcal{G}_{CC} was better than that on \mathcal{G}_{NN} . The performance of the registration was improved by determining the structure of the graphical model based on the statistical analysis of the dependencies of the variables.

5 Summary

In this article, we reported on the performance comparison study of the inference on the graphical model, which was employed for representing a statistics of the point distribution model of an organ surface. The model is registered to 3D X-ray CT images by inferring the marginal posterior distribution of the coordinates of each of the points.

For this inference on the graphical model, we can use GS and NBP. The accuracies of the estimates differ according to the inference method, even when an identical graphical model is used. The experimental results showed that the estimates obtained by GS were more accurate than those obtained by NBP. This conclusion was given significant evidences by the Wilcoxon signed rank-sum test.

It was also demonstrated that the registration performance changed according to the structure of the graphical model, even when the model was constructed based on an

Table 3. A comparison of registration errors with respect to the methods for determining the structures of graphical models

E^i (pixel)	min	average	max
\mathcal{G}_{CC}	1.60	2.71	4.94
\mathcal{G}_{NN}	1.98	3.50	5.64

identical set of the corresponding points. When the number of the edges was increased in the graphical model, the accuracy of the registration became worse when NBP was used for the inference. On the other hand, when GS is used for the inference, the accuracy did not change. These results were also supported by the Wilcoxon signed rank-sum test.

Our future works include to analyze the relationship between the metric of surfaces used for generating the corresponding points and the performance of the registration. We obtain a different statistical model according to the metric of the surfaces, even when an identical set of training surfaces is used for learning the model.

References

1. Bishop, C.M.: *Pattern Recognition and Machine Learning*. Springer, Heidelberg (2006)
2. Hontani, H., Watanabe, W.: Point-Based Non-Rigid Surface Registration with Accuracy Estimation. In: Computer Vision and Pattern Recognition, pp. 446–452 (2010)
3. Sudderth, E.B., Ihler, A.T., Isard, M., Freeman, W.T., Willsky, A.S.: Nonparametric belief propagation. *Communication of the ACM* 53, 95–103 (2010)
4. Simonson, K.M., Drescher, S.M., Tanner, F.R.: A statistics-based approach to binary image registration with uncertainty analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 112–125 (2007)
5. Murphy, K., Weiss, Y., Jordan, M.I.: Loopy Belief Propagation for Approximate Inference: An Empirical Study. In: Proceedings of Uncertainty in AI, pp. 467–475 (1999)
6. Han, T.X., Ning, H., Huang, T.S.: Efficient Nonparametric Belief Propagation with Application to Articulated Body Tracking. In: Computer Vision and Pattern Recognition, pp. 214–221 (2006)
7. Cates, J.E., Fletcher, P.T., Styner, M.A., Shenton, M.E., Whitaker, R.T.: Shape Modeling and Analysis with Entropy-Based Particle Systems. *Information Processing in Medical Imaging*, 333–345 (2007)
8. Bickel, P.J., Levina, E.: Covariance regularization by thresholding. *Ann. Statist.* 36, 2577–2604 (2008)
9. Meinshausen, N.: A Note on the Lasso for Gaussian Graphical Model Selection. *Statistics and Probability Letters* 78, 880–884 (2008)
10. Donner, R., Reiter, M., Langs, G., Peloschek, P., Bischof, H.: Fast active appearance model search using canonical correlation analysis. *IEEE Transaction on Pattern Analysis and Machine Intelligence* 28, 1690–1694 (2006)
11. Book, S., Gelman, A.: Inference and Monitoring Convergence (chapter for Gilks, Richardson, and Spiegelhalter book), vol.10 (2007)
12. Weiss, Y., Freeman, W.T.: Correctness of belief propagation in graphical models with arbitrary topology. *Neural Computation* 13, 2173–2200 (2001)

A Large-Scale Manifold Learning Approach for Brain Tumor Progression Prediction

Loc Tran¹, Deb Banerjee¹, Xiaoyan Sun¹, Jihong Wang⁴, Ashok J. Kumar⁴,
David Vinning⁴, Frederic D. McKenzie³, Yaohang Li², and Jiang Li¹

¹ Departments of ECE, ²CS, ³ MSVE, Old Dominion University, Norfolk, VA 23529

⁴ Diagnostic Imaging, University of Texas MD Anderson Cancer Center, Houston, TX 77030

Abstract. We present a novel manifold learning approach to efficiently identify low-dimensional structures, known as manifolds, embedded in large-scale, high dimensional MRI datasets for brain tumor growth prediction. The datasets consist of a series of MRI scans for three patients with tumor and progressed regions identified. We attempt to identify low dimensional manifolds for tumor, progressed and normal tissues, and most importantly, to verify if the progression manifold exists - the bridge between tumor and normal manifolds. By mapping the bridge manifold back to MRI image space, this method has the potential to predict tumor progression, thereby, greatly benefiting patient management. Preliminary results supported our hypothesis: normal and tumor manifolds are well separated in a low dimensional space and the progressed manifold is found to lie roughly between them but closer to the tumor manifold.

1 Introduction

With the rapid advancement of diagnostic imaging technology, multi-dimensional, large-scale, and heterogeneous medical datasets are generated routinely in clinical imaging exams. For instance, MR diffusion tensor imaging (DTI) has become a routine component of the brain MR imaging exams in many institutions. Although this new imaging modality together with the traditional T1, T2 or FLAIR weighted MRI scans has provided additional information and has shown potential for better brain tumor diagnosis, interpreting these large-scale, high-dimensional datasets simultaneously is challenging [1, 2]. Due to the fact that significant correlations exist among these multi-dimensional images, we hypothesize that low-dimensional geometry data structures (manifolds) are embedded in the high-dimensional space. Those manifolds might be hidden from human viewers because it is challenging for human viewers to interpret high-dimensional data. The hidden manifolds correctly extracted from the high-dimensional space may provide particularly useful information for brain cancers studies. For example, one may investigate the residence of cancer and normal tissues on the manifolds to derive rules to accurately classify cancer regions. Moreover, the bridge manifolds connecting the cancer and normal tissue manifolds may provide hints for identifying cancer progression trajectory, which can be used for predicting future tumor growth.

Many manifold learning algorithms essentially perform an eigenvector analysis on a data similarity matrix whose size is n by n , where n is the number of data samples. The memory complexity of the analysis is at least $O(n^2)$, which is not feasible for very large datasets in terms of both computational and storage requirements for a regular computer. To solve this problem, statistical sampling methods are typically used to sample a subset of data points as landmarks, a skeleton of the manifold is then identified based on the landmarks and the remaining data points can be inserted into the skeleton by a number of methods such as Nystrom approximation, column-sampling and locally linear embedding (LLE) [3-5]. To keep a faithful representation of the original manifold, effective sampling should be considered. Undersampling will distort true embedded geometry structures and thus lead to subsequent manifold learning failure while oversampling may introduce unnecessary noise. For example, the landmark MDS performs poorly for randomly chosen landmarks if the data is noisy (contains outliers) [6]. Also, data may sometimes collapse to a central point in the low dimensional space if certain "important" samples are missing [5].

Mathematical models have been studied in recent years for glioma tumor growth prediction. These models are usually classified into three major types: microscopic, mesoscopic and macroscopic. Microscopic models describe the growth process in the sub-cellular level, concentrating on activities that happen inside the tumor cell. Mesoscopic approaches focus on interactions between tumor cells and their surrounding tissue while macroscopic approaches focus on tissue level processes considering macroscopic quantities such as tumor volume and blood flow [7]. Most of the macroscopic methods use a reaction-diffusion model based on a diffusion equation introduced by Murray [8]. These models usually consist of a set of parameters which are estimated from data and have shown predictive values [9-10].

In this paper, we developed a different brain tumor progress prediction method using a data-driven manifold learning approach. We first selected a set of landmarks from a large dataset based on an importance function learned from the data, based on which we learned a manifold skeleton using the local tangent space alignment (LTSA) algorithm [11]. We then inserted the remaining data points into the skeleton using the LLE method [12]. There are several parameters to be optimized including the number of landmarks needed and the number of neighbors in the LTSA algorithm. We defined two cost functions for optimizing the parameters. We applied the method to MRI datasets from three brain tumor patients aiming to predict tumor progression by searching for the "bridge" manifold.

2 Method

2.1 Data Preparation

The MRI data of three brain tumor patients were collected using various MRI scans including FLAIR, T1-weighted, post-contrast T1-weighted, T2-weighted, and DTI. Five scalar volumes were also computed from the DTI volume including apparent

diffusion coefficient (ADC), fractional anisotropy (FA), max-, min-, and middle-eigenvalues yielding a total of ten image volumes for each visit of every patient. Each patient went through a series of visits over a time span of two years. For each patient, a rigid registration was utilized to align all volumes to the DTI volume at the first visit using the vtkCISG toolkit [13]. After registration, each pixel location can be represented by a ten dimensional feature vector corresponding to the ten MRI scans. We then selected two visits denoted as “visit 1” and “visit 2” with expanded tumor regions in visit 2 for our experiments. A radiologist defined the tumor regions on the post-contrast T1-weighted and FLAIR scans, respectively. We also defined normal regions far away from the tumor regions for training purposes. Figure 1 shows example MRI slices overlaid on the defined tumor and normal regions.

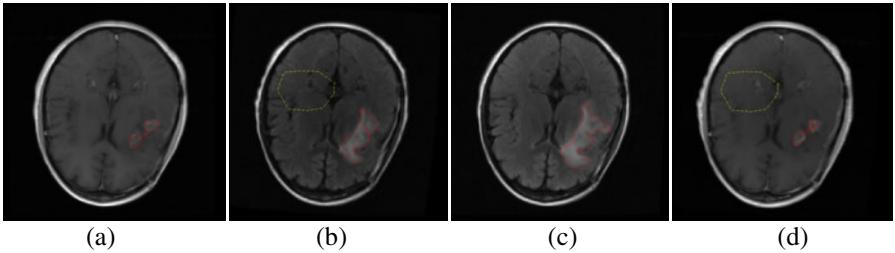


Fig. 1. Tumor and normal regions defined for patient A where the red tumor regions are labeled by a radiologist and the yellow polygon denotes normal regions. (a-b) FLAIR images at visit 1 and 2. (c-d) Post-contrast T1 images at visit 1 and 2.

2.2 Proposed System

There are roughly 65k (each slice contains 256x256 pixels) high dimensional data points in one MRI slice. We propose to use the LTSA algorithm [11] to learn the low dimensional manifolds at visit 1. Applying the LTSA algorithm to all data points is not feasible for a regular PC as we discussed above. Therefore, we developed an advanced sampling technique to select a set of landmarks based on which we learned a manifold skeleton. We then inserted the remaining data points into the skeleton using the LLE algorithm [11]. By combining the learned manifold at visit 1 with those tumor and normal regions defined at visits 1 and 2, we designed methods for predicting tumor progression and attempted to identify the “bridge” manifold, which is associated with the progressed tissues from visit 1 to visit 2. The system diagram of the proposed framework is shown in Fig. 2 and will be described as below.

Sampling based on Local Tangent Space Variation (LTV): To keep a faithful representation of the original manifold, landmarks should be carefully selected from the original data. Ideally, landmarks should be the smallest subset that can preserve the geometry in the original data. Fig. 3 shows a toy dataset to illustrate the basic

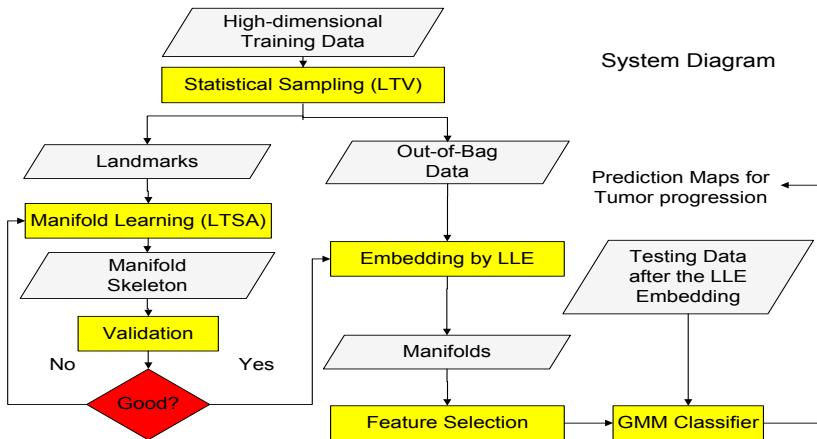


Fig. 2. System diagram for the manifold learning and tumor progression prediction

idea of LTV. Heuristically, to preserve the data structure after sampling, we should keep more data points near area ‘A’ rather than area ‘B’ in Fig. 3 because data structures near ‘A’ change more abruptly.

Based on this observation, we assigned an importance value for each of the points by computing the local tangent space variation for it. For each data point in the dataset, we found its k -nearest neighbors and performed a local principle component analysis on the k -nearest neighbors including itself. We then identified the eigenvector (spans the tangent space) corresponding to the largest eigenvalue as the red arrows shown in Fig. 3. For each data point, it has k such eigenvectors and we computed the mean value of angles between its eigenvector and the eigenvectors of all of its k -nearest neighbors. We then normalized the importance values across all data points such that they sum to one. We then sampled the dataset to obtain a set of landmarks based on the importance values.

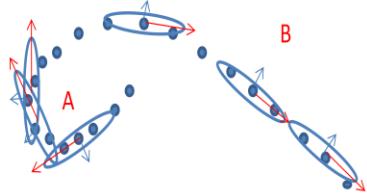


Fig. 3. The Concept of Local Tangent Space Variation

Validation of the Learned Manifold Skeleton: Successful manifold learning depends on many factors such as parameter choices for the learning algorithm and numerical stability of the algorithm. In this study, we need to determine the number of landmarks to be selected and the number of neighbors to be used by the LTSA algorithm. We proposed two cost functions as follows,

$$SF = \frac{\sum_{i,j,i \neq j} (d_M(i,j) - d_m(i,j))^2}{\sum_{i,j,i \neq j} d_M^2(i,j)} \text{ and } Acc = \frac{\text{No. of correctly classified training data}}{\text{Total no. of training data}}$$

where SF stands for the stress function and Acc represents the training accuracy after manifold learning, $d_M(i,j)$ denotes the geodesic distance between data points i and j and $d_m(i,j)$ represents their Euclidean distance. Intuitively, if the nonlinear manifold in high dimensional space is successfully unfolded, the Euclidean distance between points i and j will be the same as that of the geodesic distance in the low dimensional space. If a set of labeled training dataset is provided, which is the case in our study, the Acc value is a good criterion to verify the manifold learning. Otherwise, the stress function can be used in an unsupervised manner requiring no label information.

Embedding by LLE: Once the manifold skeleton is learned, we utilized the LLE algorithm [12, 5] to insert the remaining data points into the manifold skeleton as shown in Fig. 4, where the red dots are landmarks consisting the manifold skeleton, and the yellow square is a remaining data point to be embedded into the skeleton. We used three steps to perform this task. 1) Discovered K nearest landmarks in the original data space for the yellow square, 2) Computed a linear model that can best reconstruct the yellow square using the K landmarks and 3) inserted the yellow square into the skeleton by reusing the reconstruction weights in the linear model.

Feature Selection and Classifier Training: The learned low dimensions are usually not equally important for the subsequent classification. We ranked those dimensions by the Fisher score [14] based on the data points in the labeled regions at visit 1. In our experiments, we found that the Fisher score for features beyond the 3rd one were two orders of magnitude lower than the highest score. Next, we trained a Gaussian mixture model (GMM) using the Expectation Maximization (EM) algorithm on the labeled data. We then produced a probability map for the MRI slices at visit 1 and the probability map was used to generate a binary classification by thresholding.

3 Experiments and Results

3.1 Results for a Simulated Dataset – The Swiss Roll

Figure 5 a-d) show results for the ‘Swiss roll’ dataset. Fig. 5a) is the original dataset having 2000 data points. Fig. 5b) shows the learned results based on 900 randomly selected landmarks ($SF = 0.4527$). Fig. 5c) is the result based on 900 landmarks selected based on the LTV concept ($SF = 0.4293$). Fig. 5d) illustrates the result for a very large Swiss Roll dataset having 20k data points. A direct manifold learning for this dataset using a regular PC is prohibitive, we utilized the manifold skeleton learned in Fig. 5c) and inserted all the 20k data points into the skeleton based on the LLE algorithm. Result in (b) is slightly worse than that in (c) as expected.

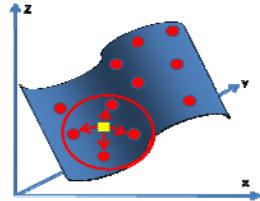


Fig. 4. Illustration of LLE embedding

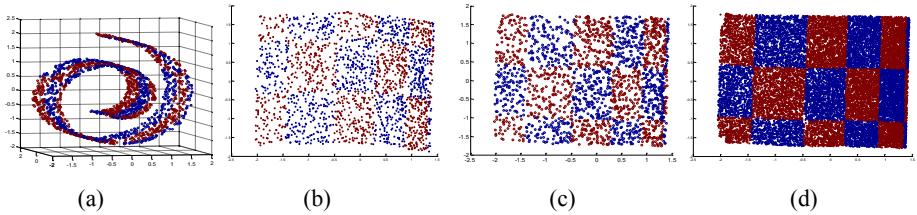


Fig. 5. Manifold learning results for swiss roll datasets

3.2 Results from the Marked FLAIR Slices at Visit 1

In this experiment, the tumor regions marked on the FLAIR slices at visit 1 were used as the ground truth for tumor and similarly sized normal regions were selected far away from the tumor regions. We first selected a set of landmarks from the marked tumor and normal regions and optimized the skeleton manifold learning based on the *Acc* criterion. Then we embedded all data points at visit 1 inside the skull into the skeleton followed by the Fisher score ranking to keep the top three dimensions. Finally, we trained a GMM model using the selected landmarks and applied the trained model to all data points inside the skull at visit 1 to obtain prediction maps. As shown in the first column in Fig. 6, the model provided a strongly localized region for the location of the tumor with noise in other regions. The brightness of a pixel corresponds to the probability from the GMM. Using a threshold of 0.5, we formed a classification mask. To remove the noise, we segmented out only the largest blob in

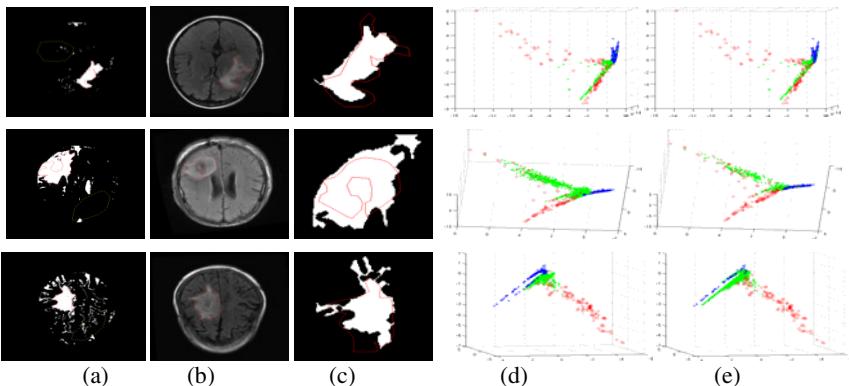


Fig. 6. Results based on the tumor regions defined on FLAIRs at visit 1. (a) GMM prediction overlaid on visit 1. (b) Original FLAIRs at visit 2. (c) Post-processed prediction results overlaid on slices at visit 2. (d) Scatter plot of the three selected dimensions, where red circles denote predicted tumor samples, blue crosses denote predicted normal samples, and green dots denote predicted tumor samples outside the tumor regions defined at visit 1 (i.e., predicted progression regions at visit 2). (e) Green dots denote actual progressed tumor samples. The progression region's ground truth was obtained by subtracting the tumor regions at visit 1 from those at visit 2. Note that only the information at visit 1 was utilized to train the prediction model.

the classification mask as shown in the third column. While the predicted tumor regions are usually beyond the marked regions, this expansion into the ambiguous area may provide the potential region of growth of the tumor. Column (b) shows the FLAIR image at visit 2 and we overlaid them on the processed binary map in column (c). It can be seen that the tumor regions at visit 2 are mostly covered by the probability map. For the low dimensional feature space plots in columns d-e, one is the predicted and another is the ground truth, there are clear separations between tumor and normal tissues (red and blue dots). The green dots in (d) denote predicted tumor tissues that fell outside of the tumor regions at visit 1. These points might be used to predict the tumor progression at visit 2 and can be considered to form a “bridge” between tumor and normal tissues. In (e), the green dots denote the actual progression points obtained by referring to the tumor regions defined at visits 1 and 2.

3.3 Results from the Marked Post-contrast T1 Slices at Visit 1

We obtained similar results in this experiment except that the predicted regions extend far beyond those defined at visits 1 and 2. This may due to the fact that the prediction may be dominated by the information in the FLAIR slices. We also computed average sensitivity and specificity for the three patients based on the defined tumor and normal regions and compared them with those without manifold learning as shown in Table 1. In the Table, sensitivities at visit 2 represent the accuracies of the predicted tumor regions based on information at visit 1 matching the tumor regions defined at visit 2, which can be interpreted as the prediction accuracy. We found that manifold learning can significantly improve the classification results.

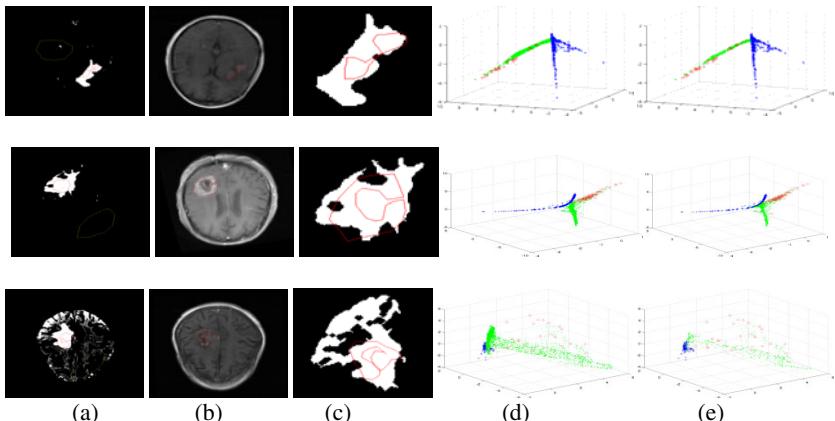


Fig. 7. Prediction results using the tumor regions defined on Post-contrast T1 slices at visit 1

4 Conclusion and Future Work

In this paper, we showed a possible nonlinear transition between normal and tumor brain tissues in high dimensional MRI datasets. Using the landmark sampling, we reduced the sample size of the MRI dataset so that conventional nonlinear dimensionality reduction techniques can be performed. There is a distinct separation

Table 1. Average sensitivities and specificities for the three patients

	Sensitivity at Visit 1	Specificity at Visit 1	Sensitivity at Visit 2
FLAIR with LTSA	0.957	1.000	0.770
FLAIR w/o LTSA	0.798	0.976	0.651
Post-contrast T1 with LTSA	0.987	0.997	0.876
Post-contrast T1 w/o LTSA	0.957	0.976	0.677

between normal and tumor tissues in the low dimensional space. We also showed that the points belonging to the tumor progression tend to accumulate between the normal and abnormal clusters. Our future work includes testing this method on more patients.

References

- [1] Pauleit, D., Langen, K.J., et al.: Can the Apparent Diffusion Coefficient be Used as A Noninvasive Parameter to Distinguish Tumor Tissue from Peritumoral Tissue in Cerebral Gliomas? *J. Magn. Reson. Imaging* (20), 758–764 (2004)
- [2] Bode, M.K., Ruohonen, J., et al.: Potential of Diffusion Imaging in Brain Tumors: A Review. *Acta Radiol.* (47), 585–594 (2006)
- [3] Deshpande, A., Rademacher, L., et al.: Matrix approximation and projective clustering via Volume Sampling. In: *Symposium on Discrete Algorithms*, vol. (2), pp. 225–247 (2006)
- [4] Drineas, P., Mahoney, M.W.: On the Nystrom Method for Approximating a Gram Matrix for Improved Kernel-Based Learning. *JMLR* (6), 2153–2175 (2005)
- [5] Wang, S., Yao, J., et al.: Improved classifier for computer-aided polyp detection in CT colonography by nonlinear dimensionality reduction. *Med. Phys.* 35(4), 1377–1386 (2008)
- [6] Silva, V., Tenenbaum, J.B.: Global Versus Local Methods in Nonlinear Dimensionality Reduction. In: *NIPS*, vol. 15, pp. 721–728 (2003)
- [7] Hatzikirou, H., Deutsch, A., et al.: Mathematical Modelling of Glioblastoma Tumor Developement: A Review. *Mathematical Models and Methods in Applied Sciences* 15(11), 1779–1794 (2005)
- [8] Murray, J.: *Mathematical Biology*. Springer, Heidelberg (1989)
- [9] Atuegwu, N.C., et al.: The integration of quantitative multi-modality imaging data into mathematical models of tumors. *Physics in Medicine and Biology* 55, 2429–2449 (2010)
- [10] Cobzas, D., Mosayebi, P., Murtha, A., Jagersand, M.: Tumor Invasion Margin on the Riemannian Space of Brain Fibers. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) *MICCAI 2009. LNCS*, vol. 5762, pp. 531–539. Springer, Heidelberg (2009)
- [11] Zhang, Z., et al.: Principal manifolds and nonlinear dimensionality reduction via local tangent space alignment. *SIAM Journal of Scientific Computing* 26(1), 313–338 (2004)
- [12] Roweis, S.T., Saul, L.K.: Nonlinear dimensionality reduction by locally linear embedding. *Science* 290(5500), 2323–2326 (2000)
- [13] Hartkens, T., Rueckert, D., et al.: VTK CISG Registration Toolkit: An open source software package for affine and non-rigid registration of single- and multimodal 3D images. In: *Workshop Bildverarbeitung fur die Medizin*, pp. 409–412 (2002)
- [14] Duda, R., Hart, P.: *Pattern Classification and Scene Analysis*, pp. 114–129 (1973)

Automated Detection of Major Thoracic Structures with a Novel Online Learning Method

Nima Tajbakhsh, Hong Wu, Wenzhe Xue, and Jianming Liang

Biomedical Informatics, Arizona State University, Scottsdale, AZ, 85259, USA
`{Nima.Tajbakhsh,Hong.Wu,Wenzhe.Xue,Jianming.Liang}@asu.edu`

Abstract. This paper presents a novel on-line learning method for automatically detecting anatomic structures in medical images. Conventional off-line learning requires collecting all representative samples before the commencement of training. Our presented approach eliminates the need for storing historical training samples and is capable of continuously enhancing its performance with new samples. We evaluate our approach with three distinct thoracic structures, demonstrating that our approach yields competing performance to the off-line approach. This demonstrated performance is attributed to our novel on-line learning structure coupled with histogram as weaker learner.

Keywords: Thoracic structure detection, on-line learning, histogram, Kalman filter.

1 Introduction

Automated detection of anatomic structures is an essential functionality in navigating large 3D image datasets and supporting computer-aided diagnosis (CAD). Several approaches have been proposed for localizing anatomic structures. Zheng et al. [7] trained a probabilistic boosting tree with steerable features to detect the heart chambers. Zhou et al. [8] combined the concepts of prediction and detection to more efficiently and precisely locate the left ventricle in ultrasound images. A framework to simultaneously locate several anatomic structures was proposed in [3]. Criminisi et al. [4] introduced context-rich visual features and utilized regression forests to estimate the organ position and the corresponding bounding box. Despite the great progress, all of these suggested methods in the literature are off-line and have to be trained in advance, offering no capability for continually improving their performance.

It would be a significant advantage if a detector could be corrected dynamically according to the feedback from physicians. This feedback is obtainable, because in the current clinical practice, all CAD systems are operating in a closed-loop environment—any CAD finding must be approved or rejected by a physician. However, the on-line methods proposed in literature (e.g., [1], [5], [6], and [9]) are basically designed for real time object tracking which demands a high level of adaptation to the object’s changes during the tracking process. Such strong adaptability is achieved at the expense of lower stability which, as

reported by Sternig et al. [5], results in a “too adaptive” detector with short memory of previous samples. However, strong long-term memory is an essential for the task of anatomy detection; therefore, such highly adaptable methods are not expected to perform well for detecting anatomic structures.

In response to this challenge, we propose a novel on-line learning algorithm which offers a satisfactory level of adaptability and better keeps the performance on the previous samples. The proposed method eliminates the need for storing historical training samples, and is capable of continuously enhancing its performance with new samples. The closest work to ours is the one developed by Grabner and Bischof [1]. Despite some similarities, our approach significantly differs from theirs [1] in both the learning structure and the adopted weak learner. The proposed learning structure is simpler and more computationally efficient. The superiority of our method is corroborated by our experiments where the proposed learning structure surpasses [1] and yields comparable results to that of off-line detector [2]. Of particular importance are how the model responds to outliers and how the model maintains the accuracy on the previous observed samples while adjusting to the new upcoming examples. Our experiments demonstrate that the proposed method achieves high stability indicating that updating the model for new samples does not degrade the model’s performance on the previous training samples. In addition, the learning method offers a satisfactory level of adaptability while providing robustness against outliers.

The proposed method is evaluated with three major thoracic structures, the pulmonary trunk (PT), carina, and aortic arch. The reason behind our choices comes from their high diagnostic value. Detecting the PT is considered as the first stage to separate pulmonary arteries from the pulmonary veins which significantly enhances the performance of CAD systems designed for pulmonary embolism detection. Carina localization is critical for our pulmonary embolism detection system, because it directly helps extract the airway. Parallel with the pulmonary artery, airway indirectly helps us remove false positives generated in the azygous vein. Finally, locating the aortic arch is fundamental to any automatic method for detecting and measuring aortic calcification.

The structure of the paper is organized as follows. Section 2 presents the proposed method, Section 3 discusses the contributions of the paper in details. Experimental results are presented in Section 4 and finally, the paper is concluded in Section 5.

2 Proposed Method

The proposed approach is an on-line feature selection method which continually updates a linear classifier. Given a set of training samples, it dynamically updates a pool containing M features and returns a subset of N best features ($N < M$) along with their corresponding voting weights.

2.1 Histogram-Based Weak Learner

Each feature in the pool is assigned one weak learner (WL) which is comprised of two histograms, one for positive samples and one for negatives. Since samples

arrive sequentially, the two histograms are built over time. In order to continually update the histograms, the range of each feature in the pool must be known in advance. To that end, the range of each feature is estimated by examining feature values computed from a temporary set of samples. This temporary set is further discarded and is not used in training stage. After approximating the ranges, they are equally divided into 100 bins. Note that, future training samples whose feature values fall out of the obtained ranges are assigned to either first or last bin. When a training sample arrives, depending on the label, one histogram is selected and the frequency of the associated bin is updated. This process is followed for every feature in the pool. In the next step, a threshold for each WL is chosen so that maximum discrimination is obtained for the pertinent feature.

2.2 Learning Algorithm

Figure 1 shows the pseudo code of the proposed on-line learning approach. When a sample arrives, all WLs are updated so as to classify the sample into the right category. If a WL manages to classify the sample, it will be rewarded by the importance weight λ which takes high values for difficult and low values for easy training samples. Note that, when it comes to choosing the first best feature for ever training sample, all WLs are rewarded or punished by $\lambda = 1$. The reward and punishment a WL receives are recorded by λ^{corr} and λ^{wrong} which are further used to calculate the error rate of each WL. Having obtained the error rate of all WLs, one can choose the best WL producing the least error rate. Then, the index of this WL is recorded in set A whose members are not considered when choosing the next best WL. Having the best WL, one can compute the voting weight $\alpha_n = \frac{1}{2} \ln\left(\frac{1 - error_{m^*}}{error_{m^*}}\right)$ which explains how the WL contributes to the final classifier. Next, the importance weight of the sample is updated . The degree to which λ is changed is proportional to the discrimination power of the best WL. Indeed, if the selected WL has already classified many samples correctly and fails to classify this sample, λ is increased significantly, on the other hand, if this WL is known to produce a poor performance, it cannot notably increase λ . The rational behind using and updating λ is to select the next best WL in a way to correct for the samples which are misclassified by previously chosen best WLs—hard training samples. Finally, $F(x) = sign\{\sum_{i \in A} \alpha_i * WL_i(x)\}$ gives the decision for each sample where WL_i is the output returned by i^{th} best WL and takes 1 if the sample contains the desired object otherwise -1.

3 Contributions

The closest sibling to our approach is [1], but our approach is dramatically different in both the learning structure and choice of WL type (Figure 2), offering the following advantages:

The first advantage of the proposed structure (Figure 2-a) is more computational efficiency. In our approach, there exists only one selector $S(\lambda)$ which performs N times on the feature pool and keeps the track of M WLs; however, in

Input:

Training samples $S = (x_t, y_t), y_t \in \{1, -1\}, s = 1, 2, \dots, T$

A feature pool containing M features

Total number of features to be selected N ($N < M$)

Total number of bins in histogram weak learners N_{bin}

Initialize:

Initialize each weak learner: $\lambda^{corr} = 1$ and $\lambda^{wrong} = 1$

Initialize the set containing selected weak learners: $A = []$

Learning process

for $t = 1, 2, \dots, T$

 Initialize the importance weight $\lambda = 1$

{Step 1: update all weak learners}

 for $m = 1, 2, \dots, M$

$WL_m = \text{Update}(x_t, y_t, WL_m)$

 end for

{Step 2: Selecting the best N features}

 for $n = 1, 2, \dots, N$

{Step 2.1: Reward and Punish all M weak learners}

 for $m = 1, 2, \dots, M$

 if $WL_m(x_t) = y_t$

$\lambda_m^{corr} = \lambda_m^{corr} + \lambda$ if correct decision

 else

$\lambda_m^{wrong} = \lambda_m^{wrong} + \lambda$ if wrong decision

 end if

$error_m = \frac{\lambda_m^{wrong}}{\lambda_m^{corr} + \lambda_m^{wrong}}$

 end for

{Step 2.2: Select the n^{th} best weak learner}

$m^* = \arg \min_{\{m \notin A\}} error_m$

 Add m^* to set A

{Step 2.3: Calculate the voting weight}

$\alpha_n = \frac{1}{2} \ln \left(\frac{1 - error_{m^*}}{error_{m^*}} \right)$

{Step 2.4: Update λ }

 if $WL_m^*(x_t) = y_t$

$\lambda = \frac{\lambda}{2(1 - error_{m^*})}$ if correct decision

 else

$\lambda = \frac{\lambda}{2(error_{m^*})}$ if wrong decision

 end if

 end for

 end for

Output

Final classifier $F(x) = sign\{\sum_{i \in A} \alpha_i * WL_i(x)\}$

Fig. 1. The pseudo code of our proposed on-line learning approach

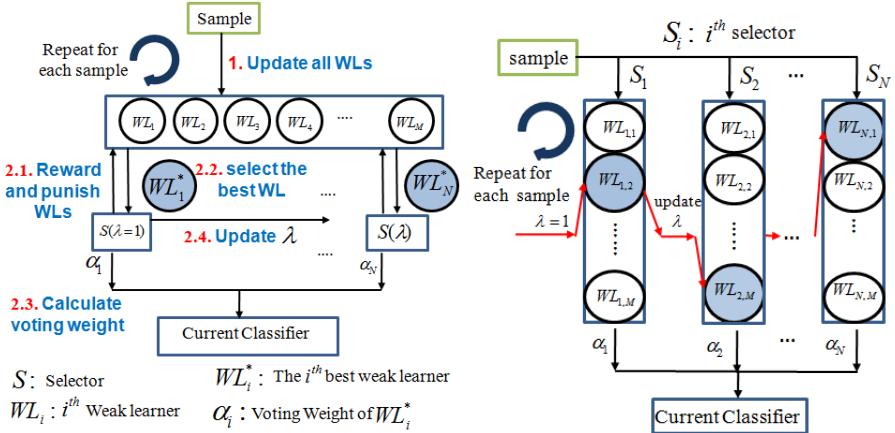


Fig. 2. (Left) An overview of our proposed approach. See the pseudo code shown in Figure 1 and follow the steps. Although our approach was inspired by the Grabner method (Right), it is dramatically different in both the learning structure and the realization of WLs.

Figure 2-b, there exist N selectors ($S_i, i = 1, 2, \dots, N$) which are to track $M \times N$ WLs over time.

The second advantage is higher resilience to outlier samples. The structure suggested in [1] offers high level of adaptability, thus potentially vulnerable to outliers. The main reason behind the observed vulnerability comes from the fact the WLs in each layer are trained and specialized for a particular range of importance weight λ , thus vulnerable to an unexpectedly large λ of an outlier sample. However, in the proposed method, WLs are not specialized for a specific range of λ , instead they are all exposed to the same set of λ values, thus not readily affected by outliers. This property can be demonstrated by the following experiment.

Assume we have trained two classifiers based on the proposed learning algorithm and the one suggested in [1]. Now we inject some outliers to the both classifiers. Outliers are very difficult samples identified by an off-line detector. When a sample arrives, both classifiers are updated and their performance are evaluated on the previously seen training samples. Then, the area under the ROC plot is computed. Following the same procedure for each new sample, one can plot the area under the curve (AUC) as a function of sample number and investigates how AUC changes over time. The more consistent the AUC, the more robustness is gained against outliers. Figure 3 shows the robustness analysis for the proposed and Grabner's method. As it is seen, both methods exhibit degradation in AUC for the first 75 samples; however, as more samples are injected, the proposed approach reveals more stability and robustness.

Note that, robustness against outliers is achieved at the expense of less adaptability. In general, there always exists a compromise between adaptability and

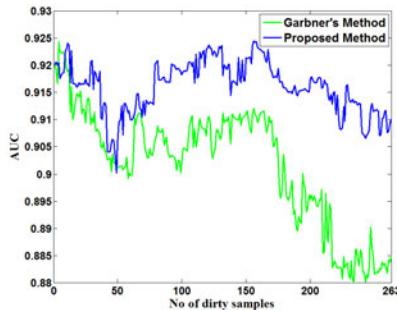


Fig. 3. Robustness analysis for the proposed and Grabner’s method. Note that, in the case of Grabner’s method, as more dirty samples are injected, less stability is observed.

stability. In the case of anatomic structure detection, this compromise should be in favor of stability which is essential for the detector to yield consistent performance over time.

The third advantage is that the proposed structure better keeps the performance on the previous samples. In [1], high level of adaptability (prompt adjustment to the new samples) is achieved at the expense of losing classification accuracy for previously observed samples. However, in contrast with Kalman filter [1], histograms can efficiently handle both past and future samples, offering high level of stability to the learned model. The other problem is that KF produces one threshold thus works the best for features with bi- model distributions. However, in practice, features need to be modeled using multi-modal distributions. Therefore, it seems essential to define multiple thresholds in order to fully utilize the existing features.

4 Experiments

The experiment exploits 157 CT pulmonary angiogram datasets, of which 80 datasets are used for training the decision model and the rest are for testing. Each dataset contains 450 ~ 600 slices. To construct the training samples, 80 training patients are scanned and extracted positive and negative samples are resized 25x25 (Figure 4). Because of simplicity and both memory and computational efficiency, 2D Haar features [2] are computed for each 25x25 training sample. Using four Haar patterns of different positions, scales, and aspect ratio, we obtain 101,400 features for each training sample. In this work, use of 2D features produces a highly efficient system with satisfactory performance. Further enhancement is expected using 3D samples and 3D Haar features which imposes much computational burden on the detection system.

Table 1 summarizes the detection rates for the proposed method and those of [2] and [1] –as the representatives of off-line and on-line detectors, respectively. A false detection is not located within 30 mm from the ground truth.

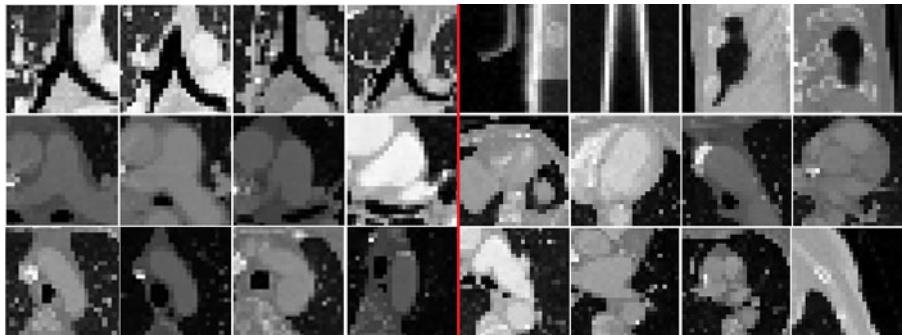


Fig. 4. (left) Four instances of the carina (first row), pulmonary trunk (second row), and aortic arch (third row). As it is seen, the samples of the aortic arch and pulmonary trunk exhibit more variation in shape and orientation. (right) 12 negative samples selected from the training patients.

Comparison with Grabner and Bischof [1]: In the case of carina, the proposed method slightly outperforms the Grabner's approach. The small variation in scale and orientation accounts for the high carina detection rates. However, the situation for the pulmonary trunk and aortic arch is quite different. Table 1 shows a drop in the detection rates of both on-line approaches though the proposed method achieves higher performance. Presence of outliers and the wide range of variation in object's scale and orientation accounts for the degradation. Dealing with object variation demands a detector with strong long-term memory. Since Grabner's approach keeps shorter memory of training samples, one can observe more performance degradation.

Comparison with off-line detector[2]: The proposed method yields detection rate very close to that of off-line in the case of the carina, and provides competitive results in the case of pulmonary trunk and aortic arch. Figure 5 compares the precision of the proposed and off-line detector. Each histogram is the distribution of distances calculated between detection points and ground truth for each anatomic structure. As it is seen, while the proposed method offers high precision, the off-line detector yields more compact distributions (sharper peaks), meaning that the detections are closer to the ground truth. We believe augmenting the current approach with multiple detector training turns the on-line method into an even stronger competitor to the off line approach.

Table 1. The detection rates for the test patients. As it is seen, the proposed approach outperforms [1] and yields comparable results to that of off-line detector [2].

Methods	Detection rate		
	PT	Aortic arch	Carina
Off-line detector [2]	93.5%	93.4%	100%
Grabner and Bischof [1]	57.1%	65.8%	94.8%
Proposed	89.6%	86.8%	98.7%

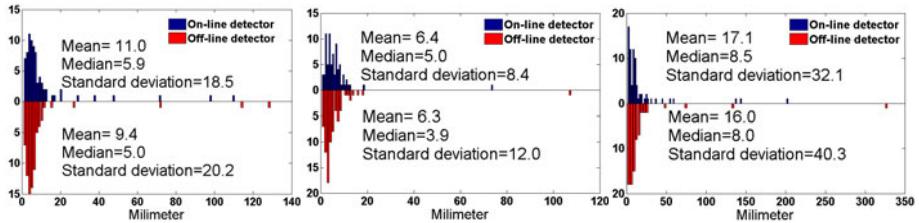


Fig. 5. (a) distribution of detection error for (left) pulmonary trunk, (middle) Carina, (right) aortic arch. "On-line detector" is trained based on the proposed learning structure with histogram WL and "Off-line detector" is the cascade structure proposed by Viola and Jones [2]. Precision of on-line and off-line detectors are comparable.

Note 1 To our knowledge, we are the first to address anatomical structure detection with on-line approaches. Other existing approaches are all off-line (e.g., [3] and [4]) with focus on detection of other major organs like heart, liver, and spleen. Accordingly, we cannot make direct comparisons with these methods.

Note 2 One might question whether progressively re-training an off-line detector eliminates the need for designing an on-line detector. Note that such a practice requires collecting and storing all previously seen and the new upcoming samples which demands a huge amount of space on the disk. In addition, re-training the model using all training samples at once (off-line) may encounter some practical issues and is limited by the amount of memory and computation time.

5 Conclusion

To our knowledge, the presented work is among the first to tackle the anatomy detection challenge from an on-line learning perspective. We proposed a new on-line learning framework and evaluated its performance for three anatomic structures with different levels of rigidity. Our experiments demonstrated that the proposed method surpasses Grabner's approach [1] for all three structures. It also yields comparable performance to off-line detector for the carina and PT. We believe the slightly poor performance for the aortic arch can be enhanced by training multiple detectors on three coronal, axial, and sagittal planes.

References

1. Grabner, H., Bischof, H.: On-line boosting and vision. In: International Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 260–267 (2006)
2. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: International Conference on Computer Vision and Pattern Recognition, vol. 1, pp. 511–518 (2001)

3. Liu, D., Zhou, K.S., Bernhardt, D., Comaniciu, D.: Search strategies for multiple landmark detection by submodular maximization. In: CVPR, San Francisco, CA, USA, pp. 2831–2838 (2010)
4. Criminisi, A., Shotton, J., Robertson, D., Konukoglu, E.: Regression forests for efficient anatomy detection and localization in CT studies. In: Menze, B., Langs, G., Tu, Z., Criminisi, A. (eds.) MICCAI 2010. LNCS, vol. 6533, pp. 106–117. Springer, Heidelberg (2011)
5. Sternig, S., Godec, M., Roth, P.M., Bischof, H.: Transientboost: On-line boosting with transient data. In: CVPRW, San Francisco, CA, USA, pp. 22–27 (2010)
6. Liu, X., Yu, T.: Gradient feature selection for online boosting. In: ICCV, Rio de Janeiro, Brazil, p. 18 (2007)
7. Zheng, Y., Barbu, A., Georgescu, B., Scheuering, M., Comaniciu, D.: Fast automatic heart chamber segmentation from 3D CT data using marginal space learning and steerable features. In: ICCV, Rio de Janeiro, Brazil, 8 pages (2007)
8. Zhou, S.K., Zhou, J., Comaniciu, D.: A boosting regression approach to medical anatomy detection. In: CVPR, Minneapolis, MN, USA, pp. 1–8 (2007)
9. Godec, M., Leistner, C., Saffari, A., Bischof, H.: On-line random naive Bayes for tracking. In: ICPR, Istanbul, Turkey, pp. 3545–3548 (2010)

Accurate Regression-Based 4D Mitral Valve Surface Reconstruction from 2D+t MRI Slices

Dime Vitanovski^{2,3,*}, Alexey Tsymbal², Razvan Ioan Ionasec¹, Michaela Schmidt³, Andreas Greiser⁴, Edgar Mueller⁴, Xiaoguang Lu¹, Gareth Funka-Lea¹, Joachim Hornegger³, and Dorin Comaniciu¹

¹ Siemens Corporate Research, Princeton, USA

² Siemens Corporate Technology, Erlangen, Germany

³ Siemens Health Care, Erlangen, Germany

⁴ Pattern Recognition Lab, Friedrich-Alexander-University, Erlangen, Germany

Abstract. Cardiac MR (CMR) imaging is increasingly accepted as the gold standard for the evaluation of cardiac anatomy, function and mass. The multi-plan ability of CMR makes it a well suited modality for evaluation of the complex anatomy of the mitral valve (MV). However, the 2D slice-based acquisition paradigm of CMR limits the 4D capabilities for precise and accurate morphological and pathological analysis due to long through-put times and protracted study. In this paper we propose a new CMR protocol for acquiring MR images for 4D MV analysis. The proposed protocol is optimized regarding the number and spatial configuration of the 2D CMR slices. Furthermore, we present a learning-based framework for patient-specific 4D MV segmentation from 2D CMR slices (sparse data). The key idea with our Regression-based Surface Reconstruction (RSR) algorithm is the use of available MV models from other imaging modalities (CT, US) to train a dynamic regression model which will then be able to infer the absent information pertinent to CMR. Extensive experiments on 200 transesophageal echocardiographic (TEE) US and 20 cardiac CT sequences are performed to train the regression model and to define the CMR acquisition protocol. With the proposed acquisition protocol, a stack of 6 parallel long-axis (LA) planes, we acquired CMR patient images and regressed 4D patient-specific MV model with an accuracy of 1.5 ± 0.2 mm and average speed of 10 sec per volume.

1 Introduction

Cardiac MR imaging emerges as the new gold standard for characterizing cardiac masses and the evaluation of cardiac function and anatomy. The multi-plan ability of CMR to acquire tomographic images in any plane, the capabilities to measure blood flow velocity in all three dimensions within a single slice and the non ionizing radiation gives it a significant advantage over other imaging modalities. Clinical studies have already proven that CMR is well suited for the evaluation of the complex anatomy of MV by comparing MV measurements

* Correspondence to dime.vitanovski.ext@siemens.com

extracted from CMR data with CT and US [1]. Nevertheless, the 2D slice-based acquisition paradigm of CMR limits the 4D anatomical and functional analysis of the heart by long throughput times and protracted study which can be utilized for more accurate and precise morphological and pathological analysis.

Although many works [2,3,4], mainly based on the seminal contribution of Cootes et al. [5] on Active Shape Models (ASM), have studied 4D chamber segmentation to overwhelm the 2D limitations of CMR, there is still no established method to extract 4D anatomical and functional information of the heart valves. Conti et al. [6] have manually initialized contours of the MV in each of the 18 radial long-axis 2D CMR slices and used interpolation to obtain a 3D model of the MV. Nevertheless, this method is characterized with a long acquisition time (18 2D slices), manual initialization, a static MV model and therefore not applicable in clinical practice. On the other hand, 4D aortic and mitral valve estimation from other imaging modalities like CT and US have been introduced by [7,8].

Within this paper we propose *a novel CMR acquisition protocol* for non-invasive assessment of the mitral valve anatomy and morphology together with a *regression-based method for patient-specific 4D MV model estimation*. Based on extensive experiments on simulated data (Section 3.1) we first defined the acquisition protocol and optimize it with respect to the number and spatial configuration of the 2D+CMR slices resulting in reduced acquisition time and 4D MV segmentation error. Second, we defined the segmentation task as a regression problem (Sec. 2.1) between the full surface and a sparse one extracted from the incomplete 2D+MRI slices. We solved the problem by learning additive boosting regression and its stabilization (Sec. 2.3) - bagging with random feature subspacing (BRFS) from MV models extracted from other imaging modalities (CT, US) used as a prior knowledge. Furthermore, experiments in Section 3.2 prove the concept of learning a regression model from one imaging modality and applying it to another one for 4D MV model estimation.

2 Methods

2.1 Regression-Based Surface Reconstruction (RSR)

In contrast to ASM-based methods, regression-based solutions make no implicit assumption about multivariate normality of the data. Comparing to simpler heart anatomies such as the left and right ventricle, the complex structure of MV exhibits higher variability and mesh point distributions different from normal. Thus, MV is more challenging to segment, especially from sparse data and requires a more robust technique.

In *regression* a solution to the following optimization problem is normally sought [9]:

$$\hat{\mathcal{R}}(\mathbf{x}) = \operatorname{argmin}_{\mathcal{R} \in \mathfrak{S}} \sum_{n=1}^N L(y(\mathbf{x}_n), \mathcal{R}(\mathbf{x}_n)) / N \quad (1)$$

where \mathfrak{S} is the set of possible regression functions, $L(\circ, \circ)$ is a loss function that penalizes the deviation of the regressor output $\mathcal{R}(\mathbf{x}_n)$ from the true output, and N is the number of available training examples. In our case the reconstruction task is defined as a regression problem between the full surface model of MV and the respective sparse data acquired using the proposed CMR protocol:

$$\mathbf{y}_{surface} = \hat{\mathcal{R}}(\mathbf{x}_{sparse}) + \epsilon \quad (2)$$

In our regression problem both for input and output data we focus on shape information and ignore respective volume data. Thus, the output $y_{surface}$ is always a set of m 3D points defining the MV surface:

$$\mathbf{y}_{surface} = ((x_1, y_1, z_1), \dots, (x_m, y_m, z_m))^T \quad (3)$$

2.2 Invariant Shape Descriptors

The input, \mathbf{x}_{sparse} , are shape descriptors (SD) describing the cloud of points belonging to MV in the sparse CMR data. The simplest but reliable solution is to use the coordinates of known points as input [10]. The obvious drawback of this solution is the necessity to provide point correspondence, which is not always feasible, especially for the data with high variability such as MV surface. A different solution, which we exploit here, is to use angles, distances and areas between random sampled points as point cloud descriptors [11]:

- **A3:** Measures the angle between three random points;
- **D2:** Measures the distance between two random points;
- **D3:** Measures the square root of the area of the triangle between three random points;

For the different shape descriptors proposed by [11] we measured feature importance by analysing the features selected by additive boosting. We have identified (A3, D2 and D3) to be most informative in our context with the average probability of occurrence 0.11, 0.07, and 0.13, correspondingly. In addition, all three types are translation, rotation and scale invariant descriptors which overcome the necessity of point correspondence. Finally, histogram bins and the four first normalized central moments describing the histogram distribution are computed from the descriptors and exploited in the regression model as input.

2.3 Ensembles of Additive Boosting Regressors

Each component m regression problem $\hat{\mathcal{R}}^m$ is solved by learning using additive boosting regression (ABR) [12]. In ABR, the weak regressors ρ_t are sequentially fit to the residuals, starting from the mean \bar{y} and proceeding with the residuals of the available set of weak regressors themselves. In ABR, the output function is assumed to take a linear form as follows [12]:

$$\hat{\mathcal{R}}(\mathbf{x}) = \sum_{t=1}^T \alpha_t \rho_t(\mathbf{x}); \rho_t(\mathbf{x}) \in \mathfrak{S} \quad (4)$$

where $\rho_t(\mathbf{x})$ is a base (weak) learner and T is the number of boosting iterations.

In the spirit of [13] and [9], we use very simple weak regressors as the base learners. These include ***simple 1D linear regression*** (SLR), ***logistic stumps*** (LS) and ***decision stumps*** (DS). For SLR, in each boosting iteration a feature which results in the smallest squared loss with linear regression is added to the pool of already selected features. Each weak learner is thus a simple linear regressor of the form ($y = \beta_1 x + \beta_0$) where x is the selected shape descriptor and y is a scalar output coordinate. LS is a simple logistic function on one shape descriptor x :

$$y = \frac{1}{1 + e^{-z}}, z = \beta_1 x + \beta_0 \quad (5)$$

Finally, DS is a piecewise linear threshold function where threshold θ is selected so that the variance in the subsets of instances produced is minimized. It is important to note that SLR results in a linear solution overall, while DS and LS seek for non-linear solutions.

Generalization performance improvement of the underlying regression models and avoidance of overfitting is achieved by injecting randomization in the input data and random features sub spacing (BRFS), similar to [14]. In particular, instead of providing a single model \mathcal{R} for the training set X , we generate a set of models \mathcal{R}_i^j , each obtained using the same additive regression procedure but on a *random sample* of the data, with instances S_i obtained using random sampling with replacement, and a subset of features F_j including 50% features randomly sampled without replacement from the original set. The final solution is then simply the mean surface for the surfaces obtained with the regressors generated from the random samples: $\mathcal{R} = \text{mean}_{i,m}(\mathcal{R}_i^m)$.

2.4 Mitral Valve Model Estimation from 2D+t CMR Slices

The mitral valve, located between the left atrium and the left ventricle, includes *posterior leaflet* composed by the posterior leaflet tip, posterior and anterior commissures and the postero annular midpoint, *anterior leaflet* defined by anterior leaflet tip, the left/right trigone and two commissures, *annulus, free edge*. Point distribution model (Fig. 1(left)) is used to represent the mitral valve surfaces S_{MV} .

Annulus and Free Edge Detection. Similar to [15] we define our joint context landmark set between the posterior annulus and the free edge landmarks (PA, PFE) and the anterior leaflet (AA, AFE), respectively (Fig.1(middle)). On each 2D LAX CMR slice we apply 2D landmark classifiers, trained with PBT [16] and 2D Haar-like features to detect the annulus and free edge landmarks independently. From the candidates generated by the detectors we select the top M candidates for the annulus detection and the top N candidates for the free edge detection. In the second stage the joint context is build from all possible candidates' pairs $\langle \text{annulus plane}, \text{free edge plane} \rangle$. In the final stage a context operator C^O is applied to compute the Haar-like features from the set of all possible candidates used to train a joint context classifier for MV landmark detection.

Mitral Valve Contour Estimation. From previously detected landmarks we initialize the contours, parameterized by 17 discrete points, as a straight line and search for edges along the normals. A least-square approach is used to fit a parametric NURBS curve to the discrete set of detected contour points C_i .

Regression-Based Full 4D MV Surface Reconstruction. In this step shape descriptors - SD are computed from the landmarks and contours detected from the 2D+t CMR slices and set into the trained regression model as input. As a result, full dynamic MV model is estimated in the 2D+t MRI slices(Fig.1(right)).

$$\mathbf{S}_{MV} = \mathcal{R}(SD[(PA, PFE, AA, AFE), C]_{1..t}) \quad (6)$$

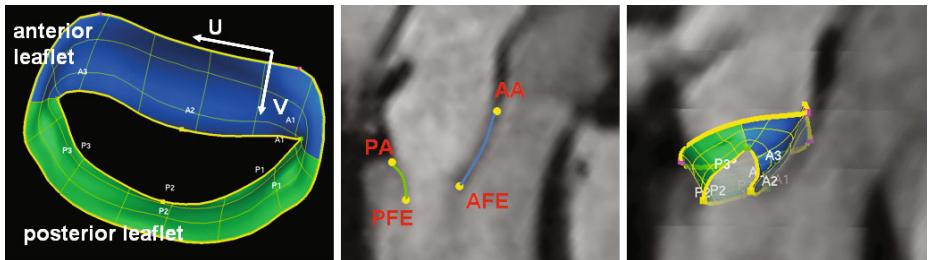


Fig. 1. *Left:* MV model. *Middle:* MV landmarks and contours. *Right:* MV model regressed into the patient-specific CMR anatomy.

3 Experimental Setting and Results

3.1 MRI Acquisition Protocol Definition

CMR scanner (1.5T) with phased-array receiver coil and breath-hold acquisition was used to acquire cine images for MV function analysis. Full cardiac cycle was covered by using a retrospectively gated ECG signal.

The mitral valve imaging plane was defined by acquiring four-chamber, three-chamber and short-axis view in the diastolic phase of the cardiac cycle. Initial orientation of the imaging plane is given by the short-axis view, where the plane normal passes through the MV commissures. Subsequently, parallel slices were defined along the normal between the commissures.

In order to find the best trade-off between MV segmentation error and acquisition time the CMR imaging protocol was defined based on simulated data. The MRI imaging plane with the same orientation was placed in 200 US TEE studies and used to obtain sparse images. To make the simulation more realistic we add gaussian noise to the plane location and orientation. The regression-based surface reconstruction method as introduced in Section 2.1 and Section 2.4 was applied to segment the MV from the simulated sparse images in the end-diastolic (ED) and end-systolic (ES) phase of the cardiac cycle.

Based on the experiments from the simulated data (see Table 1) a stack of 6 parallel LA planes results in best trade-off between MV segmentation error and acquisition time (4.86 min for a expert). A protocol with 6 radial LA planes was also considered as an option. However, due to the long acquisition planning time, the complicated plane settings and the plane mis-registration characteristic for this acquisition protocol, we found that a stack of 6 parallel planes is more appropriate for MV segmentation. **Please note that with 6 parallel 2D+t MRI slices only one third of the MV anatomy is covered.**

Table 1. MRI acquisition protocol optimization: analyzing the reconstruction error for different number of MRI imaging planes defined parallel between the MV commissures

No. Planes	2	3	4	5	6	7	8	9	10
ED (mm)	6.7 ± 1.1	5.6 ± 1.0	4.4 ± 0.9	3.5 ± 0.68	3.1 ± 1.1	2.6 ± 1.0	2.3 ± 1.0	2.1 ± 1.0	1.8 ± 0.86
ES (mm)	2.9 ± 1.2	2.6 ± 2.2	2.2 ± 1.5	2.1 ± 1.4	2.1 ± 1.2	2.3 ± 1.6	2.5 ± 2.3	2.1 ± 1.1	1.9 ± 1.0

3.2 Mitral Valve Surface Reconstruction

The proposed framework for personalized 4D MV model estimation in sparse data was evaluated on a large set of simulated data (200 US TEE and 20 cardiac CT sequences) and on *6 ECG gated CMR studies acquired according to our protocol* as introduced in Section 3.1. Each volume in the data set is associated with annotation, manually generated by experts, which is considered as ground truth. A point-to-mesh error[7] was used for the evaluation of all presented results.

Intra and inter modality accuracy of our RSR method with respect to different weak learners was evaluated. The inter modality accuracy was evaluated by training the regression model on images simulated from US TEE data and tested on simulated sparse images from CT data. For the intra modality accuracy a 3-fold cross validation was used to divide the US TEE data set into training (used to train the regression model) and test data (used to evaluate the reconstruction error).

Table 2 summarizes the patient-specific MV surface reconstruction error from incomplete data for the best CMR plane configuration protocol: stack of 6 parallel planes. We also show how different types of weak learners(simple 1D linear regression - **SLR**, logistic stump - **LS** and decision stump - **DS**) influence the reconstruction error and how this error can be reduce by incorporating all three weak learners into the framework of bagging with random feature sub-sampling (**BRFS**). Each additive boosting regression model includes 100 weak regressors with 0.2 shrinkage. For BRFS, 10 bootstrap samples of examples were generated, each with 10 random feature subsets including 50% original descriptors (resulting in 100 additive boosting regression models to combine).

With the results of the inter modality accuracy and the clinical studies which prove the compactness of the MV anatomy between different imaging modalities[1], we have shown that a regression model can be learned from one imaging modality (US) and used to estimated a patient-specific MV model in

Table 2. Left: Intra modality RSR accuracy (US). **Right:** Inter modality RSR accuracy (Train on US, reconstruct in CT).

mm	SLR	LS	DS	BRFS	mm	SLR	LS	DS	BRFS
no noise	1.7 ± 0.7	1.8 ± 0.9	2.1 ± 0.9	1.6 ± 0.4	no noise	2.2 ± 0.6	2.5 ± 0.6	2.9 ± 1.0	1.6 ± 0.6
2mm noise	1.8 ± 0.4	1.9 ± 0.7	2.3 ± 1.5	1.8 ± 0.7	2mm noise	2.6 ± 1.2	2.5 ± 0.7	3.1 ± 1.0	1.8 ± 0.8
3mm noise	2.0 ± 0.6	1.9 ± 0.5	3.0 ± 3.1	1.9 ± 0.5	3mm noise	3.0 ± 1.3	3.5 ± 1.0	3.8 ± 0.9	2.3 ± 0.5

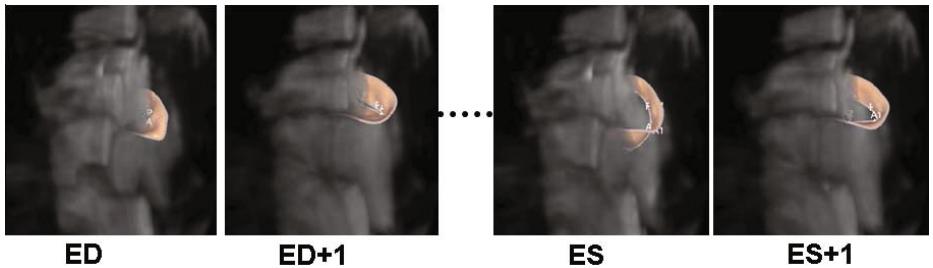


Fig. 2. Example of the reconstructed 4D MV model over the cardiac cycle from 6 parallel 2D+t MRI slices which only partially cover the MV anatomy

other imaging modality (CT). Finally, the regression model was learned offline with BRFS on all available US and CT data (1295 3D volumes) and evaluate on the CMR studies acquired according to the proposed protocol. A surface reconstruction error of 1.5 ± 0.2 mm was achieved within 10 sec per volume for the CMR studies. Figure 2 illustrates the estimated MV models for the ED and ES phase of the cardiac cycle for the CMR studies.

3.3 Results on Regression-Based Surface Reconstruction vs ASM

As benchmarks for our regression-based reconstruction we consider two ASM-based techniques, which we call ASM1 and ASM2 in our experiments. Both solutions start with generating a PCA-based ASM for the set of training data which includes the complete set of shape coordinates and descriptors. Then, at the segmentation phase, ASM1, is a least squares fitting to the incomplete set of available normalized coordinates and shape descriptors. Analytically this solution is obtained using only rows corresponding to the available descriptors in the eigenvector matrix and the provided features themselves [Draper and Smith, 1998]. ASM2 is a hill-climbing search in the space of PC mode values so as to minimize MAE (mean absolute error) of the reconstructed surface with respect to the available coordinates and descriptors. Starting from the first principal component, the mode values are iteratively refined so as to reproduce the available incomplete data (point coordinates and shape descriptors) as close as possible. As the focus here is on reconstruction quality and not on robustness to noise (the main input data are validated and rather noise-free), we do not consider

other advanced ASM-based variants here, which attempt to regularize the solution and reduce overfitting. A surface reconstruction error of $2.8 \pm 0.8\text{mm}$ and $3.1 \pm 1.6\text{mm}$ was achieved for ASM1 and ASM2, respectively.

4 Conclusion

This paper presents a novel CMR acquisition protocol for fast non-invasive 4D anatomy and function assessment of the mitral valve. A regression-based surface reconstruction (RSR) method, designed according to the protocol, is used to learn from the existing MV models in US and CT data and applied to estimate patient-specific 4D MV model from sparse CMR data. With our protocol and method we overcame some of the CMR limitations: long through-put time, protracted study, 2D function and anatomy analysis. Furthermore, a 4D model of the MV can be utilized to extract more accurate morphological and pathological information over the cardiac cycle. Extensive experiments on simulated (CT and US) data have proven our concept of learning a regression model from one modality (US) and applying it on other one (CT) for 4D MV surface reconstruction. Following this concept we learn the RSR algorithm on all available CT and US data and demonstrate a reconstruction accuracy of $1.5 \pm 0.2\text{mm}$ within 10 sec pro volume for the CMR studies.

References

1. Djavidani, B., et al.: Planimetry of mitral valve stenosis by magnetic resonance imaging. In: American College of Cardiology, vol. 45, pp. 2048–2053 (2005)
2. van Assen, H.C., et al.: Spasm: a 3d-asm for segmentation of sparse and arbitrarily oriented cardiac mri data. In: Medical Image Analysis, vol. 10, pp. 286–303 (2006)
3. Frangi, A.F., et al.: Threedimensional modeling for functional analysis of cardiac images: A review. IEEE Trans. Medical Imaging 20, 2–25 (2001)
4. Wang, X., et al.: Reconstruction of detailed left ventricle motion from tmri using deformable models. In: Functional Imaging and Modeling of the Heart (2007)
5. Cootes, T.F., et al.: Active shape models-their training and application. Computer Vision and Image Understanding 61, 38–59 (1995)
6. Conti, C.A., et al.: Mitral valve modelling in ischemic patients: Finite element analysis from cardiac magnetic resonance imaginge. In: Computing in Cardiology, pp. 1059–1062 (2010)
7. Ionasec, R., et al.: Patient-specific modeling and quantification of the aortic and mitral valves from 4d cardiac ct and tee. In: TMI (2010) (in Press)
8. Grbić, S., Ionasec, R., Vitanovski, D., Voigt, I., Wang, Y., Georgescu, B., Navab, N., Comaniciu, D.: Complete valvular heart apparatus model from 4D cardiac CT. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6361, pp. 218–226. Springer, Heidelberg (2010)
9. Zhou, S.K., et al.: Image based regression using boosting method. In: ICCV (2005)
10. Vitanovski, D., Tsymbal, A., Ionasec, R.I., Georgescu, B., Huber, M., Taylor, A., Schievano, S., Zhou, S.K., Hornegger, J., Comaniciu, D.: Cross-modality assessment and planning for pulmonary trunk treatment using CT and MRI imaging. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6361, pp. 460–467. Springer, Heidelberg (2010)

11. Osada, R., et al.: Shape distributions. *ACM Transactions on Graphics* 21, 807–832 (2002)
12. Friedman, J.H.: Greedy function approximation: A gradient boosting machine. *Annals of Statistics* 29, 1189–1232 (2000)
13. Viola, P., Jones, M.: Rapid object detection using a boosted cascade of simple features. In: *CVPR* (2001)
14. Webb, G.I.: Multiboosting: A technique for combining boosting and wagging. *Machine Learning*, 159–196 (2000)
15. Lu, X., et al.: Discriminative joint context for automatic landmark set detection from a single cardiac mr long axis slice. In: *FIHM* (2009)
16. Tu, Z.: Probabilistic boosting-tree: Learning discriminativemethods for classification, recognition, and clustering. In: *ICCV* (2005)

Tree Structured Model of Skin Lesion Growth Pattern via Color Based Cluster Analysis

Sina KhakAbi^{1,2,3}, Tim K. Lee^{1,2,3}, and M. Stella Atkins^{1,2}

¹ School of Computing Science, Simon Fraser University, Canada
`{skhakabi,stella}@sfu.ca`

² Department of Dermatology and Skin Science, University of British Columbia and Vancouver Coastal Health Research Institute, Canada

³ Cancer Control Research Program, BC Cancer Research Centre, Canada
`tlee@bccrc.ca`

Abstract. This paper presents a novel approach to analysis and classification of skin lesions based on their growth pattern. Our method constructs a tree structure for every lesion by repeatedly subdividing the image into sub-images using color based clustering. In this method, segmentation which is a challenging task is not required. The obtained multi-scale tree structure provides a framework that allows us to extract a variety of features, based on the appearance of the tree structure or sub-images corresponding to nodes of the tree. Preliminary features (the number of nodes, leaves, and depth of the tree, and 9 compactness indices of the dark spots represented by the sub-images associated with each node of the tree) are used to train a supervised learning algorithm. Results show the strength of the method in classifying lesions into malignant and benign classes. We achieved Precision of 0.855, Recall of 0.849, and F-measure of 0.834 using 3-layer perceptron and Precision of 0.829, Recall of 0.832, and F-measure of 0.817 using AdaBoost on a dataset containing 112 malignant and 298 benign lesion dermoscopic images.

1 Introduction

Malignant melanoma is one of the most frequent types of cancer in the world. In recent decades, the annual rate of its incidence has been increasing by 3%-7% in fair-skinned populations [12]. It is increasing faster than any other cancers in the world, and in 2008, melanoma was the sixth most common malignancy in men and the seventh in women [6]. Despite the fact of lethality of skin cancer, if the malignant lesion is detected early, it can be cured without complication. Hence, there is a growing demand for the computer-aided-diagnosis of melanoma to improve the diagnostic accuracy.

There are several different modalities to obtain images of skin lesions. One reliable way to screen skin lesions is to use a dermoscope [4], a non-invasive hand-held device which captures magnified digital skin images using either polarized light or oil immersion to render the outermost layer of the skin, called

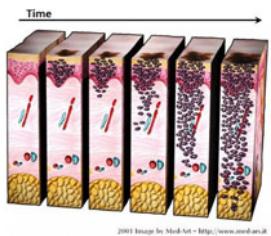


Fig. 1. The vertical growth phase of pigmented skin lesion over time. Melanocytes penetrate into dermis and change the coloration of the lesion. (©2001 by Med-Art <http://www.med-ars.it>)

the epidermis, translucent. This enables dermatologists to recognize the subsurface structures that are associated with malignancy. Therefore, image processing algorithms can be incorporated to facilitate the diagnosis.

For the conventional use of dermoscopy without computer assistance, dermatologists have proposed different scoring methods that facilitate the diagnostic process. The following methods are mainly used for this purpose [2]: 1)*ABCD rule*; 2)*pattern analysis*; 3)*Menzie's method*; 4)*7-point checklist*; and 5)*texture analysis*. Every method highlights specific features which are biologically inspired, arising from the structure of malignant lesions. Features extracted using these methods are mainly based on shape structure, texture and color variation. Recently, many studies have been carried out for extracting the diagnostic features using image processing techniques. In [11], an overview of some of these techniques is provided.

In this study, a novel approach inspired by the analysis of the growth pattern of skin lesions is proposed. A malignant lesion is often identified with two growth phases, radial and vertical [5]. Both malignant and benign lesions start growing radially. In this phase, pigmented lesion is formed by nests of melanocytes, which synthesize a brown pigmentation called melanin. This phase happens in the epidermis. Invasive melanoma forms in the vertical phase when malignant melanocytes start penetrating into dermis (see Fig. 1). In this phase, an invasive melanoma tends to show multiple colors due to the position of melanin in different skin layers, formation of blood vessels, and regression of the lesion. Melanin is dark brown in epidermis, tan near the dermoepidermal junction, and blue-gray in the dermis. This is caused by different absorbance of light in different layers of skin. Accordingly, pigmented malignant and benign lesions demonstrate a radial growth pattern which can be observed using dermoscopy. This pattern is also used in [13] as a spatial constraint to achieve a better segmentation of the lesion.

We use a multi-scale tree structure for analyzing the growth pattern of lesions. This framework allows us to study: *i*) radial growth pattern of the individual dark spots, and *ii*) distribution of the dark spots over the lesion. In section 2, the method for constructing the tree structure is explained. This structure provides a strong framework to extract various features. In section 3, some preliminary features are discussed to demonstrate the strength of the provided framework. Section 4 describes the implementation in detail. Section 5 presents performance evaluation of the method based on the features extracted in section 3. Section 6 concludes the paper.

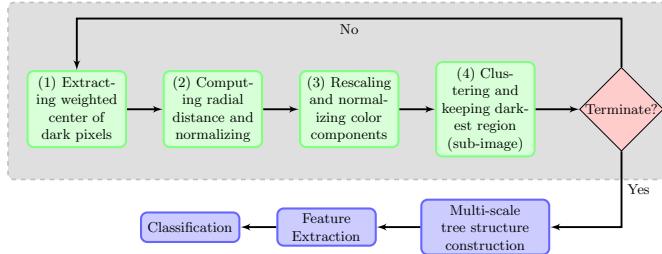


Fig. 2. Flowchart of the provided algorithm

2 Method

We study two main features based on the coloration of the lesion. Our novel method takes a top-down approach in which every skin lesion image is repeatedly sub-divided into sub-images through a clustering procedure over the pixels. At every iteration of the sub-division, the color features and a spatial coordinate feature (see sub-section 2.2) of all pixels in the sub-image are rescaled and clustered into 2 clusters. The darkest cluster of pixels is retained and formed into a new sub-image. The process of rescaling and reclustering is repeated for the new sub-images until termination. In other words, we are constructing a multi-scale tree structure such that every sub-image corresponds to a node in the tree and its children are obtained by rescaling the image features, reclustering pixels, and retaining the darker cluster. Fig. 2 demonstrates the schematic view of our algorithm and Fig. 3 illustrates the results of applying the algorithm.

As a preprocessing step, we removed dark hairs using Dullrazor [9]. The following sub-sections explain the algorithm in detail towards creation of the multi-scale tree structure. Feature extraction and classification steps are discussed later.

2.1 Extraction of Weighted Center of Dark Pixels

For every iteration, RGB image is converted to gray-scale image using the following equation

$$I = 0.2989 \times R + 0.5870 \times G + 0.1140 \times B \quad (1)$$

then I is normalized between 0 and 1 and darker pixels with $I \leq I_d$ are extracted. As a result, X and Y which are two sets of coordinates, are defined as follows

$$X = \{x | I_{x,y} \leq I_d\} \text{ and } Y = \{y | I_{x,y} \leq I_d\} \quad (2)$$

Then, (x_C, y_C) the coordinates of the center point are calculated as follows

$$x_C = \frac{1}{|X|} \times \sum_{i=1}^{|X|} x_i \text{ and } y_C = \frac{1}{|Y|} \times \sum_{i=1}^{|Y|} y_i \quad (3)$$

Simply, x_C and y_C are the centroid of the dark pixels.

2.2 Computing Radial Distance

The radial distance $D_{x,y}$ is computed for every pixel in the current sub-image by calculating the Euclidean distance between pixels and the extracted center, where

$$D_{x,y} = \sqrt{(x - x_C)^2 + (y - y_C)^2} \quad (4)$$

Consequently, all the pixels on the same circle centered on the extracted center point, will have the same radial distance. The goal is to include this distance as a feature in the clustering stage in next step. This will lead to layered clusters of pixels which represent the growth pattern of the spot centered on the point that is the weighted average over the dark pixels.

2.3 Clustering and Shrinking

Incorporating the radial distance from the center of an image as a spatial constraint in clustering pixels is provided in [13] to facilitate skin lesion segmentation. We modified and extended this idea. Instead of clustering around the center of an image, we cluster around the centroid of a dark spot. In addition, we extend the method to multi-scale tree structure. In every iteration, we have a sub-set of pixels of the original image. This sub-set is clustered into two clusters using the well-known *k-means* clustering algorithm. In the clustering stage, a four-dimensional normalized feature set, containing three color components and the previously computed radial distance, is fed to the algorithm with different weights assigned. We chose Hue-Saturation-Intensity (HSI) color space to build the feature set. Therefore, RGB color space is converted to HSI color space using equations given in [7]. Thus, for pixel (x, y) the feature set is $\{H_{x,y}, S_{x,y}, I_{x,y}, D_{x,y}\}$. After obtaining two clusters, the darker cluster is characterized and kept by looking for the highest I value in the resulting clusters' color-map. The new sub-set of pixels in the dark cluster delimits regions over the image. These regions undergo hole filling and opening morphological operations and define our final sub-images. Every sub-image obtained so far will be entered in the loop independently and all aforementioned three steps will be repeated for them and for their descendant sub-images. This is where the idea of the tree structure arises.

Sub-image decomposition ends when one of these four conditions is true: *i*) number of pixels in the obtained sub-image is less than a constant value, *ii*) color variance over the blue and green intensity channels in a sub-image is less than a threshold, *iii*) descendant of a sub-image has not significantly changed in comparison to its parent, or *iv*) the depth of the implicitly constructed tree is more than a limit. Further details are provided in section 4.

2.4 Multi-scale Tree Structure Construction

The above process can be interpreted as a depth first search (DFS) over a tree. When a sub-image breaks into multiple sub-images, each sub-image is traversed

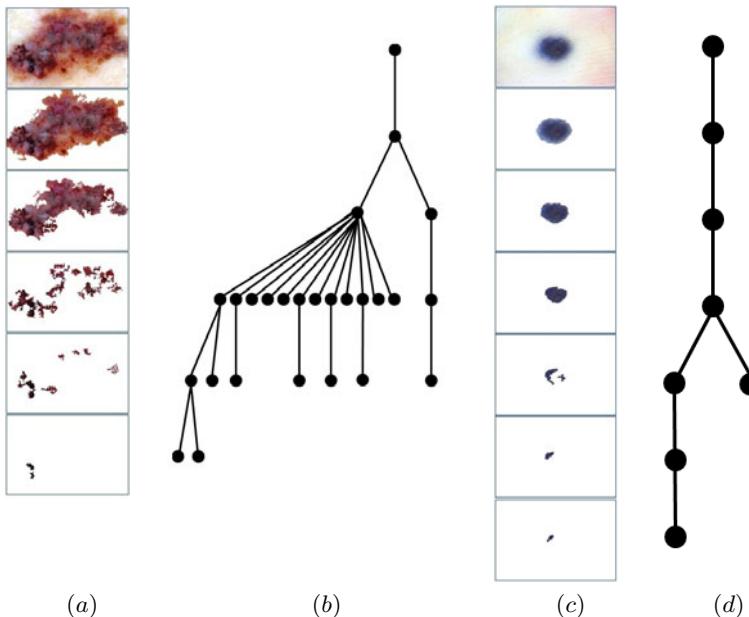


Fig. 3. (a) and (c) illustrate decomposition of a malignant and benign lesions respectively. Their corresponding tree structures are also shown in (b) and (d). Number of branches and leaves in the benign lesion is less than that of malignant lesion.

until termination, then the next sub-image is traversed and so on. Hence, the main image corresponds to the root of the tree; its sub-images form its children in the second layer; the third layer consists of sub-images of the second layer; and so forth. Fig. 3 illustrates samples of the extracted tree structure, for a benign and a malignant lesion.

Our hypothesis is that if the complexity of the tree structure is high, the probability of malignancy is also high. Complexity can be defined as the number of nodes and the degree at every node. This can be seen in examples provided in Fig. 3.

3 Feature Extraction

The tree structure framework allows us to extract a variety of features, based on the tree itself or the sub-images at each node of the tree. In this study some preliminary features are examined to illustrate the strength of the framework. The number of nodes and leaves in the tree and depth of the tree are very basic features that can be used in a classification procedure. As can be seen in Fig. 3, these features reflect the complexity of the tree structure to some extent. Compactness index is chosen as a sub-image based feature which reflects the irregularity of the border of the sub-image. It is calculated using as follows

$$CI = \frac{P^2}{4\pi A} \quad (5)$$

where P is the perimeter of the object and A is the object area. The extracted compactness index is stored in an array, where element i of the array corresponds to depth $i + 1$ of the tree. The CI is calculated for all the nodes in depth i and is summed up and multiplied by $\frac{10-i}{9}$ which is proportional to the depth. Borders of sub-images which are placed close to the root of the tree are more similar to the border of the lesion in the main image. As a result, they are more reliable and their corresponding nodes should be given more weight than nodes close to leaves. The results of using these features separately are discussed in section 5. The CI is not calculated for the root as it is the main image with square border. Thus, the length of the array should be one less than the maximum possible depth of the tree which is defined in the loop termination conditions. Other features such as border irregularity index [10] can be extracted and used instead of CI ; however, their time cost is higher than calculating CI .

4 Implementation Detail

We implemented the algorithm using the MATLAB's image processing toolbox. In this study, thresholds and parameters of the algorithm are experimentally chosen, as the algorithm is not very sensitive to these values. Extracting the center point as the first step in every iteration of the loop, is not sensitive to the threshold chosen for the gray-scale image intensities. This threshold (I_d) is set to 0.25. Increasing this value will cause incorporating of more pixels over the image and change the center point slightly. This change will affect the value of $D_{x,y}$. As the number of clusters in the clustering stage is set to two, we are assured that the darker cluster includes all the dark regions that we are interested in. Consequently, the final tree structure will have the same shape. This is an advantage of our algorithm.

The weights assigned to the H, S, and I are set to 2 and the weight of radial distance D is set to 1 in the clustering stage. Higher weighting on D results in rounded borders of sub-images in the nodes close to the leaves. The low weighting on D and higher weighting on color components reduce the sensitivity to the threshold value for center extraction step.

Values of parameters and thresholds for termination conditions are as follows: *i)* area of sub-images should be more than 500 pixels, *ii)* the variance over the both blue and green channels should be greater than 40 when the blue and green channel values are normalized between 0 and 255, *iii)* difference in area between two consecutive nodes should be more than 50 pixels, and *iv)* the maximum allowed depth is set to 10.

5 Results

We prepared a dataset of 410 images taken from [1]. These images were picked out of 763 images to provide sufficient malignant lesions in the data set for machine learning algorithms. This was done without visual inspection and only by looking at the diagnosis of lesions. In this dataset, we have 112 malignant lesion images (containing melanoma and basal cell carcinoma (BCC)), and 298 benign lesion images (containing congenital, compound, dermal, Clark, spitz, and blue nevus; dermatofibroma; and seborrheic keratosis). Our ground truth is based on the information provided in [1]. This information is based on dermatologists' diagnosis. WEKA [8] is used to classify the images. 3-layer perceptron and AdaBoost are chosen as classifiers. The parameters for 3-layer perceptron are set as follows: *learning rate* is set to 0.3, *momentum* is set to 0.2, *training time* is set to 500 and *validation threshold* is set to 20. The parameters of AdaBoost are set as follows: the *number of iterations* is set to 10, the *seed* is randomly generated and the *weight threshold* is set to 100.

All the features which are computed in the feature extraction section are gathered in a 12-dimensional feature set (number of nodes, number of leaves, depth, and 9 CI components) and the resulting set is fed into the classifiers. Validation method is set to leave-one-out ten-fold cross validation. The malignant and benign images are randomly chosen and uniformly distributed over the folds. Table 1 illustrates the classification results between malignant and benign classes for our dataset using two different classifiers. In this table, we also provide the result of the classification using Betta *et al.*'s streaks detection technique [3], which we reimplemented in MATLAB and evaluated using the same dataset. In Betta's method, the image is subdivided into 16 equal rectangular blocks and, like our method, the dark brown color is examined in *HSI* color space, though Betta uses it for detecting streaks. However, Betta's method is very sensitive to the segmentation of the lesion.

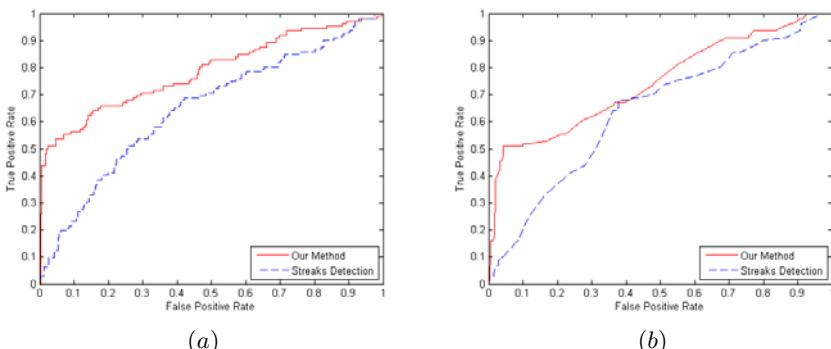


Fig. 4. Receiver Operating Characteristic (ROC) curve of classification based on our method and streaks detection method [3] using (a) 3-layer preceptron classifier and (b) AdaBoost classifier. The areas under the curves are provided in Table 1.

Table 1. Table of results of classifying the dataset into malignant and benign classes using 12 features

Method	Classifier	Precision	Recall	F-Measure	AUC of ROC
Our Method	3-layer perceptron	0.855	0.849	0.834	0.786
Streaks Detection	3-layer perceptron	0.689	0.732	0.656	0.648
Our Method	AdaBoost	0.829	0.832	0.817	0.745
Streaks Detection	AdaBoost	0.639	0.724	0.619	0.642

Table 2. Results of classifying the data-set using CI in different layers of the tree

Feature set	Classifier	Precision	Recall	F-Measure	AUC of ROC
CI1 to CI3	3-layer perceptron	0.639	0.712	0.641	0.617
CI4 to CI9	3-layer perceptron	0.713	0.729	0.622	0.494
CI1 to CI3	AdaBoost	0.692	0.732	0.685	0.637
CI4 to CI9	AdaBoost	0.596	0.722	0.614	0.490

The values reported for Precision, Recall, and F-measure are weighted average values based on the size of the class. Fig. 4 represents the ROC curve of the classifications for two methods and two different classifiers. Our method outperformed streaks detection method using both classifiers as shown in Table 1 and Fig. 4.

We also studied the effect of CI alone in lower and higher nodes in the extracted tree. Table 2 provides the results of this study using the AdaBoost and the 3-layer preceptron classifier and justifies using higher weights for nodes close to root.

6 Conclusion

We have developed a multi-scale tree structure based iterative framework for feature extraction and classification of the skin lesions. The strength of this framework is that it does not require an explicit lesion segmentation, which could be a challenging problem. As illustrated in this paper, we build a structure based on the distribution of dark spots; however, we could also build a tree structure with light color spots or regions with special characteristics and textures. We also demonstrated that a good classification result was achieved using just a few preliminary features.

We will further expand the features to include color, texture and other shape information. The optimal parameters of the algorithm could be learnt using an optimization procedure. Finally, we will validate our method on other large data sets.

Acknowledgements. This research was funded by NSERC and CIHR. The authors would also like to thank to Dr. Harvey Lui and Dr. David McLean for their guidance.

References

1. Argenziano, G., Soyer, H.P., et al.: Interactive Atlas of Dermoscopy (Book and CD-ROM). Edra Medical Publishing and New Media (2000)
2. Argenziano, G., Soyer, H.P., et al.: Dermoscopy of pigmented skin lesions: Results of a consensus meeting via the internet. *Journal of the American Academy of Dermatology* 48(5), 679–693 (2003)
3. Betta, G., Di Leo, G., Fabbrocini, G., Paolillo, A., Scalvenzi, M.: Automated application of the 7-point checklist diagnosis method for skin lesions: Estimation of chromatic and shape parameters. In: IEEE Instrumentation and Measurement Technology Conference, vol. 3, pp. 1818–1822 (2005)
4. Binder, M., Schwarz, M., et al.: Epiluminescence microscopy: A useful tool for the diagnosis of pigmented skin lesions for formally trained dermatologists. *Arch. Dermatol.* 131(3), 286–291 (1995)
5. Clark, W.H., Ainsworth, A.M., Bernardino, E.A., Yang, C.H., Mihm, C.M., Reed, R.J.: The developmental biology of primary human malignant melanomas. *Semin. Oncol.* 2(1), 83–103 (1975)
6. Erickson, C., Driscoll, M.S.: Melanoma epidemic: Facts and controversies. *Clinics in Dermatology* 28(3), 281–286 (2010)
7. Gonzalez, R.C., Woods, R.E.: Digital Image Processing, 2nd edn. Addison-Wesley Longman Publishing Co., Inc., Amsterdam (2001)
8. Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., Witten, I.H.: The weka data mining software: an update. *SIGKDD Explor. Newsl.* 11, 10–18 (2009)
9. Lee, T., Ng, V., Gallagher, R., Coldman, A., McLean, D.: Dullrazor ®: A software approach to hair removal from images. *Computers in Biology and Medicine* 27(6), 533–543 (1997)
10. Lee, T.K., McLean, D.I., Atkins, M.S.: Irregularity index: A new border irregularity measure for cutaneous melanocytic lesions. *Medical Image Analysis* 7(1), 47–64 (2003)
11. Maglogiannis, I., Doukas, C.: Overview of advanced computer vision systems for skin lesions characterization. *IEEE Transactions on Information Technology in Biomedicine* 13(5), 721–733 (2009)
12. Marks, R.: Epidemiology of melanoma. *Clinical and Experimental Dermatology* 25(6), 459–463 (2000)
13. Zhou, H., Chen, M., Zou, L., Gass, R., Ferris, L., Drogowski, L., Rehg, J.: Spatially constrained segmentation of dermoscopy images. In: 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro, ISBI 2008, pp. 800–803 (May 2008)

Subject-Specific Cardiac Segmentation Based on Reinforcement Learning with Shape Instantiation

Lichao Wang, Su-Lin Lee, Robert Merrifield, and Guang-Zhong Yang

The Hamlyn Centre for Robotic Surgery, Imperial College London, UK
`{lichao.wang, su-lin.lee, rdm99, g.z.yang}@imperial.ac.uk`

Abstract. Subject-specific segmentation for medical images plays a critical role in translating medical image computing techniques to routine clinical practice. Many current segmentation methods, however, are still focused on building general models, and thus lack the generalizability for unseen, particularly pathological data. In this paper, a reinforcement learning algorithm is proposed to integrate specific human expert behavior for segmenting subject-specific data. The algorithm uses a generic two-layer reinforcement learning framework and incorporates shape instantiation to constrain the target shape geometrically. The learning occurs in the background when the user segments the image in real-time, thus eliminating the need to prepare subject-specific training data. Detailed validation of the algorithm on hypertrophic cardiomyopathy (HCM) data-sets demonstrates improved segmentation accuracy, reduced user-input, and thus the potential clinical value of the proposed algorithm.

1 Introduction

Subject-specific medical image segmentation is critical in clinical applications of algorithm designs. Over the years, commonly used segmentation algorithms can be divided into two broad categories: methods relying mainly on the information from the current image instance, e.g., active contours and the level-set method [1, 2], and methods that incorporate prior knowledge from training data, e.g., statistical shape models (SSMs) [3]. Due to the difficulty of catering for a diverse range of appearance variations and artifacts, active contours and the level-set method can easily fail. However, these methods are readily generalizable to unseen patient data sets. SSMs, on the other hand, although proven to be robust against artifacts and noise, have the problem of over-fitting [4] and limited generalizability. For cardiac segmentation of patients with hypertrophic cardiomyopathy (HCM), for example, different myocardial segments can have significant changes in thickness, leading to much more complicated LV endocardium [5], which is difficult to be captured by SSMs. In such cases, the SSMs' tendency to constrain the shapes towards the training data causes the derived contours to deviate away from the actual target borders. In practice, the incorporation of patient data during SSM training is difficult, as the patient anatomy tends to vary considerably between subjects. A number of algorithms have been proposed to increase the generalization capability of the SSMs. These include, for example, the use of wavelet transforms as the underlying variables to alleviate the over-fitting problem in active shape models (ASMs) [6]. In [7], synthetic models were combined with

statistical modes of variation to allow for additional flexibility during segmentation. However, these methods can lead to erroneous localization results due to the lack of subject-specific information.

In clinical applications, it is known that experienced human experts are always able to adapt shape delineation to local information, regardless of whether the data is normal or pathological, or whether the image quality is high or low. They know exactly when to use the image appearance information and when to use the geometrical constraint. There have been methods proposed utilising user interaction to aid automatic segmentation. For example, the authors in [8, 9] developed different algorithms to use user input as a prior to constrain SSM fitting. Without knowledge accumulation, however, these methods might require the user to constantly perform similar corrections for different images.

In this paper, a semi-automatic algorithm that incorporates the knowledge through real-time background learning is proposed. The algorithm is based on a two-layer reinforcement learning segmentation framework [10]. Another important element of this work is the use of the principle of shape instantiation [11] to provide the geometrical constraint. The proposed algorithm is able to learn the user-specific behavior *in-situ* about how to adaptively use different strategies given different contextual information, thereby building a model readily generalizable to subject-specific data. The method is validated by segmenting a HCM data set, to quantitatively assess the segmentation accuracy as well as the reduction of the required user input compared to fully manual segmentation.

2 Methods

2.1 Two-Layer Reinforcement Learning

The proposed semi-automatic algorithm is illustrated in Fig. 1. For simplicity and the purpose of illustrating the learning framework, the algorithm uses a radial search approach commonly adopted in cardiac segmentation. From user defined contour points, weightings are learned for different segmentation strategies using the two-layer reinforcement learning algorithm [10]. To this end, two categories of segmentation strategies are needed, i.e., the appearance strategies to extract the target using the image information and the geometrical strategy, which is the shape instantiation in this paper. Through interaction with the user, the reinforcement learning agent learns how to balance the use of different strategies when the context is different. The localization of a contour point along a radial is determined by the following cost function:

$$c(\mathbf{x}, \alpha, \mathbf{w}) = \alpha \left(\sum_{i=1}^n w_i c_{A_i}(\mathbf{x}) \right) + (1 - \alpha) c_G(\mathbf{x}) \quad (1)$$

where \mathbf{x} is the pixel position on the radial, $c_{A_i}(\mathbf{x})$ is the cost of the i th strategy within the appearance strategy category, w_i is the weighting for the i th appearance strategy (i.e., the i th in-category weighting), α is the weighting for the appearance strategy category (i.e., the cross-category weighting), $c_G(\mathbf{x})$ is the cost of the geometrical constraint and \mathbf{w} is the in-category weighting vector. During the learning process, the agent makes a segmentation based on its current knowledge of the weightings by

localizing contour points along the radial directions. The user then either approves the segmentation, or corrects the shape by providing an example point on the correct contour. From the user examples, a policy gradient algorithm is carried out on the corresponding radial direction to estimate the optimal weightings of the different strategies for the in-category appearance strategy weightings:

$$\mathbf{w}_{t+1} = \mathbf{w}_t + \gamma \left| (r_{in}(\mathbf{w}))_t \right| \nabla_{\mathbf{w}} (r_{in})_t (\mathbf{w}_t) \quad (2)$$

as well as the cross-category weightings:

$$\alpha_{t+1} = \alpha_t + \gamma \left| (r_{cross}(\alpha))_t \right| \nabla_{\alpha} (r_{cross})_t (\alpha_t) \quad (3)$$

where γ is a constant coefficient, t is the time step, r_{in} and r_{cross} are the rewards of the agent's localization using only appearance strategies and all strategies respectively, which can be defined as a function of the distance between its localization and the user example on the corresponding radial direction. For the cross-category weighting estimation, an iterative policy gradient is used: the estimated weightings on the new example radial are used to update those of the previous example radials, which are then used in turn to update the weightings of the latest example again, until the estimated weightings for all the examples in the current image do not change.

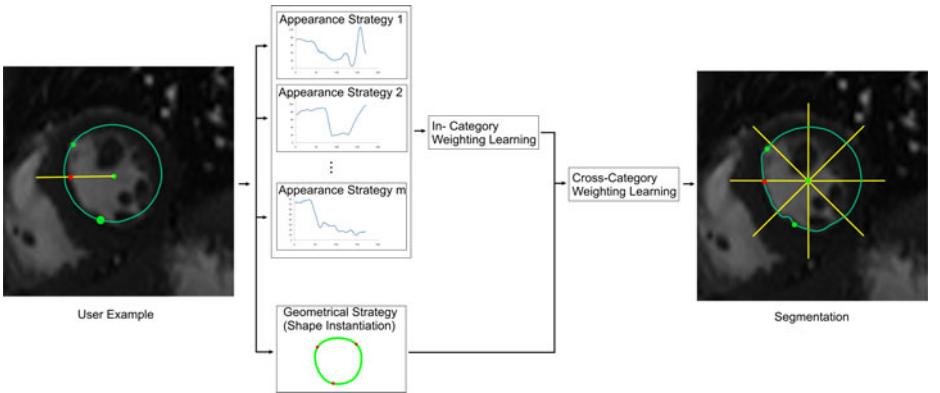


Fig. 1. The basic principle of the proposed algorithm. The two-layer reinforcement learning algorithm integrates multiple appearance strategies, as well as the shape instantiation as the geometrical localization strategy, where the first layer learns how to choose different appearance strategies, while the second layer learns to combine appearance strategies and the geometrical constraint. In this figure, the green points in the images are anchor points while the red point is the user-provided shape correction.

An important part of the two-layer reinforcement learning algorithm is the features describing the contextual information. The features can be application-specifically designed. For the cardiac LV segmentation in this paper, the features include the appearance along the radial, the radial's angular position relative to the anterior RV/LV junction point, and the smoothness of the local contour piece. The appearance along the radial direction describes the quality of the edgeness in the local area, i.e., how

likely there is a clear myocardium border on this radial. The angle information describes the likelihood of the existence of papillary muscles on the radial from the anatomical point of the view, while the smoothness feature gives the learning agent awareness of the shape's smoothness. In the space spanned by these features, the learned weightings on example radials are interpolated to all radials in the same image where there are no user examples available using Kriging [12].

2.2 Shape Instantiation for Subject-Specific Geometrical Constraint

The geometrical constraint used in the two-layer reinforcement learning algorithm is from shape instantiation, a concept of using limited sample data in conjunction with prior data to estimate the entire geometry of a dynamic shape [11]. There are many instantiation algorithms, one example of which is using regression, e.g., partial least squares regression (PLSR). PLSR decomposes both the independent variable matrix \mathbf{X} (also called the predictors) and the dependent variable matrix \mathbf{Y} simultaneously and then extracts components that are the most representative for the relationship between \mathbf{X} and \mathbf{Y} while eliminating irrelevant noise. Mathematically, the simultaneous decomposition of \mathbf{X} and \mathbf{Y} can be described as follows:

$$\hat{\mathbf{Y}} = \mathbf{T}\mathbf{D}\mathbf{C}^T = \mathbf{X}(\mathbf{P}^T)^{-1} \quad (4)$$

where $\hat{\mathbf{Y}}$ is the estimate of \mathbf{Y} , \mathbf{T} is the score matrix, \mathbf{D} is the regression weighting matrix, \mathbf{C} is the loading matrix on \mathbf{Y} and \mathbf{P} is the loading matrix on \mathbf{X} . With the estimated model, an output vector \mathbf{y} can be estimated given an input \mathbf{x} . PLSR has been used for examining the contractility of the myocardium [13] and predicting cardiac motion from surface intensity traces of the chest [14]. The capability of PLSR to effectively capture the relations between control points and the entire shape, especially those between neighboring points, makes it able to instantiate unseen shapes.

For the examples illustrated in this paper, the user defined points, i.e., two RV/LV junctions in each short-axis slice as the initial anchor points as well as those used to correct the current segmentation in the reinforcement learning process, are used as the independent variable \mathbf{x} to predict the shape \mathbf{y} . With equation (4), the shape can be estimated from the trained PLSR model, which provides a subject-specific geometrical constraint as a candidate strategy for the two-layer reinforcement learning. At the beginning of the segmentation, the initial PLSR model is trained using a set of normal LV data. After each HCM shape is segmented during the learning process, the result is added into the training set to increase the variability of the shape instantiation model.

2.3 Validation

The proposed algorithm is validated using a magnetic resonance (MR) HCM LV dataset of 21 subjects acquired from a 1.5T scanner (Siemens Sonata 1.5T, Erlangen, Germany) with a trueFISP imaging sequence (in-plane pixel resolution = 1.5 - 2 mm, slice thickness = 10 mm). To derive the ground truth data, manual delineation was performed by an expert using CMRtools (Cardiovascular Imaging Solutions Ltd., London, UK). For the initial shape instantiation model, 30 normal subjects were used (scanned with the same protocol, in-plane pixel resolution = 1.1 - 2 mm, slice thickness = 7 mm). All data were equally re-sampled both spatially and temporally.

3 Results

The proposed algorithm is applied to segment the endocardial borders of the 21 HCM cine data for the end-diastolic and the end-systolic frames. For these data, significant morphological variation was evident due to the thickened LV wall (see Fig. 3 for examples). To initialize the reinforcement learning process, the user is required to delineate three anchor points, i.e., the anterior and inferior RV/LV junctions as well as the centre of the LV blood pool. The interactive process then starts with the learning agent segmenting the image while the user corrects or approves the segmentation result. In order to measure the learning capability of the proposed algorithm, the user's input during the interactive process is based on a previously manually segmented shape, which is used as the ground truth. The appearance strategies used in this paper include three popularly used detectors for cardiac borders: the maximal derivative along the radial, a 1D intensity profile and a 1D gradient profile.

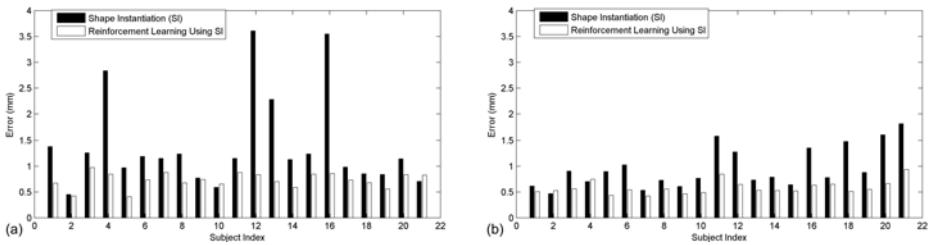
Table 1. The accuracy (in mm) and the user input required by the proposed method and manual segmentation using CMRtools for segmenting the end-diastolic (ED) and the end-systolic (ES) frames

Frame & Location	Proposed Method		Manual Segmentation		Improvement
	Error	Avg. No. of User Points	Inter-Observer Variability of Accuracy	Avg. No. of User Points	
ED-Base	0.78±0.24	3.9	1.03±0.65	8.0	51.3%
ED-Mid	0.72±0.23	3.9	0.82±0.26	8.3	53.0%
ED-Apex	0.69±0.15	4.4	0.88±0.43	6.7	34.3%
ED-All	0.73±0.15	24.4	0.91±0.37	47.2	48.3%
ES-Base	0.57±0.17	4.3	0.89±0.60	9.0	52.2%
ES-Mid	0.59±0.16	4.6	0.93±0.62	8.1	43.2%
ES-Apex	0.58±0.19	4.0	0.74±0.36	7.6	47.4%
ES-All	0.58±0.13	25.9	0.85±0.30	49.6	47.8%

The segmentation result, including the contour localization accuracy and the user interaction amount, is summarized in Table 1 for the end-diastolic and the end-systolic frames, respectively. For comparison, manual delineation of the same data was performed twice by using CMRtools to generate the intra-observer result. From Table 1, it can be seen that very high segmentation accuracy as well as consistency is achieved by the proposed algorithm, whilst the amount of user interaction is reduced significantly, for both the end-diastolic and the end-systolic frames. A paired t-test is performed to assess the statistical significance of the reduced user input. It is observed that the reduction is significant for both cases, with $p \ll 0.005$. It is also worth noting that the accuracy of the proposed algorithm is higher than the intra-observer variability. This is due to the learning capability of the proposed algorithm to acquire specific user behaviors. Additional error metrics are shown in Table 2 demonstrating the Jaccard index and the maximal radial difference of the proposed method.

Table 2. The Jaccard index and the maximal radial difference of the proposed method

Location	End-Diastole		End-Systole	
	Jaccard Index	Max. Radial Diff.	Jaccard Index	Max. Radial Diff.
Base	0.930±0.023	2.13 mm	0.909±0.019	2.10 mm
Mid	0.933±0.018	2.50 mm	0.870±0.075	2.29 mm
Apex	0.925±0.025	2.00 mm	0.862±0.060	1.78 mm
All	0.930±0.016	2.50 mm	0.892±0.031	2.29 mm

**Fig. 2.** Detailed results of the proposed method for the HCM data studied: shape instantiation (SI) and reinforcement learning with SI for end-diastole (a) and end-systole (b)

To demonstrate the use of shape instantiation as the geometrical constraint in the reinforcement learning algorithm, Fig. 2 (a) and (b) compare the 3D shape localization results using only shape instantiation based on the user points and those using the proposed algorithm with shape instantiation incorporated. It can be seen from the figures that despite the complicated morphology of the HCM LV, the incrementally updated shape instantiation model is able to capture the geometry well in most cases with the user points due to PLSR's effective extraction of the relationships between control points and the entire shape, therefore providing an efficient subject-specific geometrical constraint. The proposed algorithm, on the other hand, is able to balance the use of different appearance and geometrical strategies, thus leading to more accurate and consistent results. Even for extremely complicated cases where the shape instantiation alone produces results far away from the desired location, the proposed algorithm is able to secure the final segmentation due to the adaptive combination of all strategies. Fig. 3 provides some examples, where the results of the proposed method are compared with those only using spline interpolation based on user points, as used in CMRtools. It can be seen that even though the target shapes deviate considerably from normal LV endocardium shapes, the proposed method manages to produce accurate segmentations with a limited amount of user input. By contrast, the method that only uses spline interpolation needs a lot more user input to produce similar results. Fig. 4 shows three example error-plots of the 3D results. In the figures, the outermost ring corresponds to the basal slice while the innermost corresponds to the apical slice. The horizontal line corresponds to the anterior RV/LV junction. With the angle increasing clockwise, the plot shows errors in the order of the anterior, lateral, inferior and septal LV walls. High accuracy can be seen achieved for all regions of the 3D shapes.

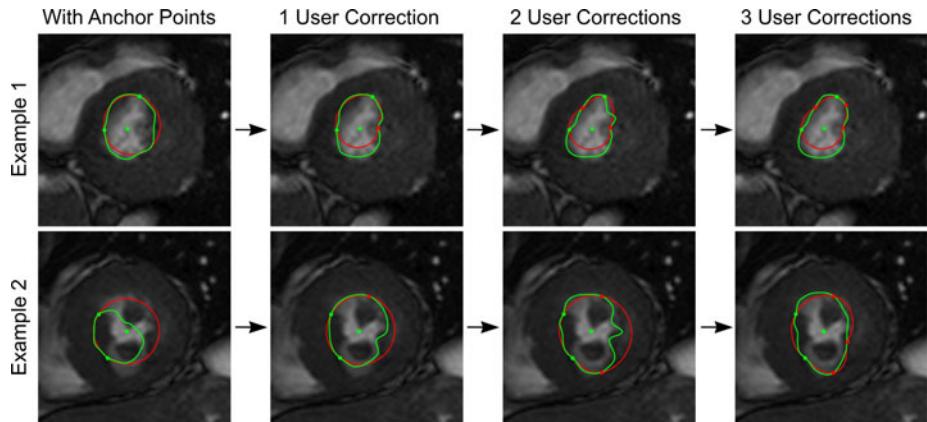


Fig. 3. Examples of semi-automatic segmentation for the HCM LV. Each row shows the consecutive steps of segmenting one 2D image. Green contour: the result of the proposed algorithm; Red contour: the result of pure spline interpolation based on user points as used in CMRTools. Green points: anchor points; Red points: example points provided by the user to correct the machine's segmentation.

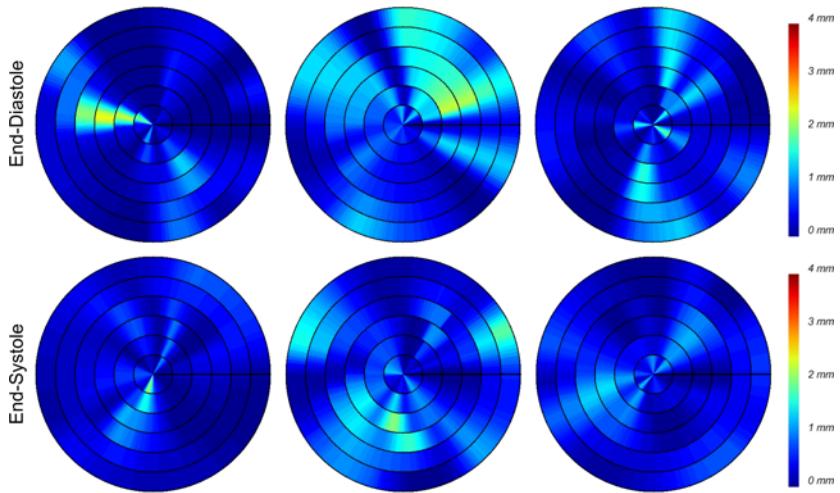


Fig. 4. Three instances of the 3D localization error using the proposed reinforcement learning algorithm

4 Conclusion

This paper introduces a semi-automatic algorithm that pervasively learns the user behavior for subject-specific cardiac segmentation. With shape instantiation as the geometric-al constraint, the reinforcement learning algorithm is able to adaptively use different

segmentation strategies according to different contextual information. Detailed validation shows the generalizability of the proposed algorithm to pathological data with considerably complicated geometry, as well as a significant reduction in user input, thus demonstrating its potential clinical value.

References

1. Xu, C., Prince, J.L.: Snakes, shapes, and gradient vector flow. *IEEE Transactions on Image Processing* 7, 359–369 (1998)
2. Chan, T.F., Vese, L.A.: Active contours without edges. *IEEE Transactions on Image Processing* 10, 266–277 (2001)
3. Cootes, T.F., Hill, A., Taylor, C.J., Haslam, J.: The Use of Active Shape Models for Locating Structures in Medical Images. In: *Proceedings of the 13th International Conference on Information Processing in Medical Imaging*, pp. 33–47. Springer, Heidelberg (1993)
4. Lekadir, K., Keenan, N.G., Pennell, D.J., Yang, G.-Z.: An Inter-Landmark Approach to 4-D Shape Extraction and Interpretation: Application to Myocardial Motion Assessment in MRI. *IEEE Transactions on Medical Imaging* 30, 52–68 (2011)
5. Cecchi, F., Yacoub, M.H., Olivotto, I.: Hypertrophic cardiomyopathy in the community: why we should care. *Nat. Clin. Pract. Cardiovasc. Med.* 2, 324–325 (2005)
6. Davatzikos, C., Tao, X., Shen, D.: Hierarchical active shape models, using the wavelet transform. *IEEE Transactions on Medical Imaging* 22, 414–423 (2003)
7. Wang, Y., Staib, L.H.: Boundary finding with prior shape and smoothness models. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 22, 738–743 (2000)
8. de Brujne, M., van Ginneken, B., Viergever, M.A., Niessen, W.J.: Interactive segmentation of abdominal aortic aneurysms in CTA images. *Medical Image Analysis* 8, 127–138 (2004)
9. Hug, J., Brechbühler, C., Székely, G.: Model-Based Initialisation for Segmentation. In: Vernon, D. (ed.) *ECCV 2000*. LNCS, vol. 1843, pp. 290–306. Springer, Heidelberg (2000)
10. Wang, L., Merrifield, R., Yang, G.-Z.: Reinforcement Learning for Context Aware Segmentation. In: Fichtinger, G., Martel, A., Peters, T. (eds.) *MICCAI 2011*, Part III. LNCS, vol. 6893, pp. 627–634. Springer, Heidelberg (2011)
11. Lee, S.-L., Chung, A., Lerotic, M., Hawkins, M.A., Tait, D., Yang, G.-Z.: Dynamic shape instantiation for intra-operative guidance. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) *MICCAI 2010*, Part I. LNCS, vol. 6361, pp. 69–76. Springer, Heidelberg (2010)
12. Matheron, G.: Principles of geostatistics. *Economic Geology* 58, 1246–1266 (1963)
13. Lee, S.-L., Wu, Q., Huntbatch, A., Yang, G.-Z.: Predictive K-PLSR myocardial contractility modeling with phase contrast MR velocity mapping. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) *MICCAI 2007*, Part II. LNCS, vol. 4792, pp. 866–873. Springer, Heidelberg (2007)
14. Ablitt, N.A., Jianxin, G., Keegan, J., Stegger, L., Firmin, D.N., Yang, G.-Z.: Predictive cardiac motion modeling and correction with partial least squares regression. *IEEE Transactions on Medical Imaging* 23, 1315–1324 (2004)

Faster Segmentation Algorithm for Optical Coherence Tomography Images with Guaranteed Smoothness*

Lei Xu¹, Branislav Stojkovic¹, Hu Ding¹, Qi Song²,
Xiaodong Wu², Milan Sonka², and Jinhui Xu¹

¹ Department of Computer Science and Engineering
State University of New York at Buffalo Buffalo, NY 14260, USA

{lxu,bs65,huding,jinhui}@buffalo.edu

² Department of Electrical and Computer Engineering
University of Iowa Iowa City, IA 52242, USA
{qi-song,xiaodong-wu,milan-sonka}@uiowa.edu

Abstract. This paper considers the problem of segmenting an accurate and smooth surface from 3D volumetric images. Despite extensive studies in the past, the segmentation problem remains challenging in medical imaging, and becomes even harder in highly noisy and edge-weak images. In this paper we present a highly efficient graph-theoretical approach for segmenting a surface from 3D OCT images. Our approach adopts an objective function that combines the weight and the smoothness of the surface so that the resulting segmentation achieves global optimality and smoothness simultaneously. Based on a volumetric graph representation of the 3D images that incorporates curvature information, our approach first generates a set of 2D local optimal segmentations, and then iteratively improves the solution by fast local computation at regions where significant improvement can be achieved. It can be shown that our approach monotonically improves the quality of solution and converges rather quickly to the global optimal solution. To evaluate the convergence and performance of our method, we test it on both artificial data sets and a set of 14 3D OCT images. Our experiments suggest that the proposed method yields optimal (or almost optimal) solutions in 3 to 5 iterations. Comparing to the existing approaches, our method has a much improved running time, yields almost the same global optimality but with much better smoothness, which makes it especially suitable for segmenting highly noisy images. Our approach can be easily generalized to multi-surface detection.

* The research of the first three and the last authors was supported in part by NSF through a CAREER Award CCF-0546509 and a grant IIS-0713489. The research of the other three authors was supported in part by the NSF grants CCF-0844765 and the NIH grants K25 CA123112 and R01 EB004640.

1 Introduction

In this paper, we consider the problem of segmenting an accurate and smooth surface in 3D volumetric images. Efficient detection of globally optimal surface in volumetric images is a central problem in many medical image analysis applications. The problem arises not only in the segmentation problem of biomedical images (e.g., CT, MRI, Ultrasound, Microscopy, Optical Coherence Tomography (OCT)) [4,5], but also in many other fundamental optimization problems.

Many known segmentation methods used today are energy-based approaches, which can be classified into two categories, variational segmentation and combinatorial segmentation. Variational methods include snake, region competition, geodesic active contours, active appearance models, and other methods based on level-sets. In most cases, variational segmentation uses variational optimization techniques that find only a local minima of the corresponding energy function. Some versions of active contours compute a globally optimal solution, but only in 2-D. The combinatorial segmentation methods are mainly based on graph algorithms, such as minimum spanning tree [3], shortest paths and their 3D extensions [1], and graph cuts [2]. Each of these approaches has respective strengths and weakness, depending on the optimality of the cost function. In recent years, Wu et al. proposed a new detection method for optimal surfaces [8]. Their optimality is controlled by a cost function using only the voxel weights and some geometric constraints on the connectivity of the surface. Their approach can compute a globally optimal surface in polynomial time. However, since their cost function only considers the weight of the voxels, the resulting surface often contains spikes and jaggedness in noisy regions. To fix this problem, a new segmentation model which emphasizes both the global optimality and local smoothness was proposed in [9]. Their approach uses an upper-bounded (mean) curvature as a penalty to avoid spikes and achieves global optimality and smoothness simultaneously. Despite many advantages of this method, there are several places that desire further improvement. For example, (1) the resulting surface is optimal in the sense that the maximum curvature is upper bounded by a pre-specified value; (2) during iterations, some portions of the surface may not change much, and thus does not need to use the same amount of effort to optimize them.

To provide faster and better solution, we study in this paper a new segmentation model, called *Curvature-UnBounded Optimal Smooth Surface (CuBOSS)*. In the CuBOSS problem, the objective function is a linear combination of voxel weights and their mean curvatures. However, instead of bounding the maximum curvature to guarantee the surface continuity, a parameter N representing the maximum allowed change in the z coordinate of the surface along the unit distance in the y direction is used as the connectivity constraint. Different from the traditional graph based algorithms using smooth parameter of both x and y directions, our approach could achieve the same smoothness with less restrictions by adjusting the parameter of the objective function. The additional requirement on the smoothness dramatically increases the hardness of the problem. There are evidences suggesting that the CuBOSS problem is NP-hard.

For the CuBOSS problem, we present a novel graph-theoretical approach. Our approach first uses 2D shortest paths to compute a set of local segmentations, and then stitches them to form a global surface. The obtained surface are then iteratively improved through local computation, with each iteration strictly reducing the objective value. We tested our approach using both artificial data sets and a set of 14 OCT images. Our experiments show that within a small constant (3-5) number of iterations, our approach converges rather quickly to optimal or almost optimal. Compared to existing approaches [9] using OCT images, our approach removes the maximum curvature restriction and avoids to perform the computation on the stable (or almost stable) portions of the surface during iterations which leads to significant improvement on the running time. Our segmentation has a total weight which is almost equal to that of the weight-only approach in ([5]), but with much improved smoothness. In all our experiments, our approach almost eliminates all spikes, seemingly suggests that our approach could handle highly noisy images.

Since all computations in our approach are based on 2D local information. Our approach is particularly suitable for fully parallel implementation on GPUs, which could potentially allow us to segment 3 or 4D volumetric images in real time.

2 Problem Description

2.1 Curvature and Mean Curvature

Curvature captures the intrinsic geometric properties of surfaces and has been widely used for measuring smoothness [6] and [7]. Recently, a new way of using mean curvature (second derivative) as a regularization factor for 3D surface is proposed in [9]. In the CuBOSS problem, we adopt the mean curvature since it is relatively easier to approximate by two orthogonal directions. In the CuBOSS problem, we use the curvatures along the x and y directions to approximate the mean curvature, which gives us the freedom to de-couple the curvature computation in the two directions. We call the curvatures in the two direction as x-curvature and y-curvature respectively. Each x and y curvature is the curvature of a 2D curve. For a 2D curve given explicitly as a function of $y = f(x)$, its curvature is $\kappa = \frac{|y''|}{(1+y'^2)^{3/2}}$. For a more general curve with parametrization $s(t) = (x(t), y(t))$, its curvature at $s(t)$ is $\kappa(t) = \frac{x'(t)y''(t)-y'(t)x''(t)}{(x'(t)^2+y'(t)^2)^{3/2}}$.

2.2 Single Surface Detection

The CuBOSS problem can be defined as follows: Given a 3D image I (see Fig. 1) of size $n_x \times n_y \times n_z$ and with each voxel $I(x, y, z)$ associated with a non-negative weight $w(x, y, z)$ representing the inverse of the probability that $I(x, y, z)$ is a boundary voxel, find a terrain surface S

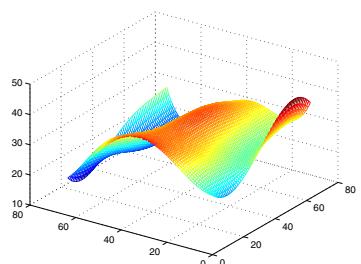


Fig. 1. Orientation of surface

(i.e., each column of I intersects S at one voxel) with the minimum cost, where the cost of S is defined as $c(S) = \sum_{I(x,y,z) \in S} \alpha w(x, y, z) + (1 - \alpha)|\kappa(x, y, z)|$, where $\kappa(x, y, z)$ is the mean curvature of S at $I(x, y, z)$, $\alpha \in [0, 1]$ is a weighting constant, and the voxel weight is computed by some edge detectors. Edge detectors are operators approximating derivatives of the image function using differences, which could be computed from one or several convolution masks. Computing the weight of the OCT data sets examined in this paper was described in [5] and [4]. An optimal surface is the one with the minimum total sum of voxel costs among all feasible surfaces that can be defined in the 3D image. The feasibility of a surface is constrained by the smoothness parameter N which is the maximum allowed change in the z coordinate of the surface along the unit distance in the y (or x) direction. Note we only restrict on one direction.

3 The Algorithm

To solve the CuBOSS problem, we propose a graph-theoretical based iterative approach. In this approach, our main idea is to detect the surface with global optimum (including smoothness) by iterative local improvement that can be converted to a sequence of shortest path problem in 2D spaces. More specifically, the algorithm for solving the CuBOSS problem contains two main steps. First, a feasible surface S is constructed as the initial solution. Second, the surface S is iteratively improved to global optimum by local adjustment. We call the first step as *initialization* and second step as *stitching*.

Initialization of CuBOSS. In the initialization step, our goal is to construct a feasible surface S with less cost. To achieve this, instead of expanding the surface construction from the first x -slice to the last x -slice one by one in [9], our strategy is to construct S of each x -slice $I(i, :, :)$ independently since there is no restriction on x -curvature. By doing this, there are two advantages, (1) the running time of each x -slice is reduced from $O(n_y n_z \kappa_{\max}^3)$ to $O(n_y n_z N^2)$ where κ_{\max} is the maximum curvature allowed of the surface defined in [9] and (2) Construction S of each x -slice independently is suitable for fully parallel implementation. To construct $S(1, :, :)$, the curve which is the intersection between the first x -slice $I(1, :, :)$ and S , our idea is to construct a weighted directed graph $G_1 = (V_1, E_1)$ of $I(1, :, :)$ and reduce it to a shortest path problem in G_1 which could be solved optimally. G_1 is constructed according to $I(1, :, :)$ as follows. Every vertex $V(u, c, b) \in V_1$ is a 3-tuple of voxels $u = I(x_u, y_u, z_u)$, $c = I(x_c, y_c, z_c)$ and $b = I(x_b, y_b, z_b)$ in $I(1, :, :)$, where the three voxels satisfy the following three conditions,

$$\begin{cases} x_u = x_c = x_b = 1, & (\text{c1: The first } x\text{-slice}) \\ y_u = y_c + 1 \text{ and } y_b = y_c - 1, & (\text{c2: 3 consecutive neighboring columns}) \\ |z_u - z_c| \leq N \text{ and } |z_u - z_b| \leq N & (\text{c3: Guarantee smoothness}) \end{cases}$$

Let $\kappa[V(u, c, b)]$ be the curvature of voxel c when voxel u and voxel b as its y -direction neighbors. We have $\kappa[V(u, c, b)] = \frac{1}{2}(\kappa_x[V(u, c, b)] + \kappa_y[V(u, c, b)])$

where $\kappa_x[V(u, c, b)]$ and $\kappa_y[V(u, c, b)]$ denote the x-direction and y-direction curvature of voxel c respectively. Since we do not consider x-direction curvature in the initialization step, $\kappa[V(u, c, b)] = \frac{1}{2}\kappa_y[V(u, c, b)]$. The cost assigned to $V(u, c, b)$ is

$$\text{cost}[V(u, c, b)] = \alpha w(x_c, y_c, z_c) + (1 - \alpha)\kappa[V(u, c, b)] \quad (1)$$

An edge is added for a pair of vertices $V(u, c, b)$ and $V(u', c', b')$ if $c = u'$ and $b = c'$ indicating that the pair of vertices share two common voxels (see Figure 2).

After adding two dummy vertices in G_1 (one as source and the other as sink), the problem in the first x-slice $I(1, :, :)$ can be reduced to a shortest path problem in G_1 . Since G_1 is a DAG (directed acyclic graph), it could be solved optimally by topological sort algorithm in linear time $O(n_y n_z N^2)$. The cost of the resulting path includes the weight and the y-curvature of each voxel on the path. Thus $S(1, :, :)$ is formed by the voxels appeared on the path. The above procedure is repeated to each rest of x-slices.

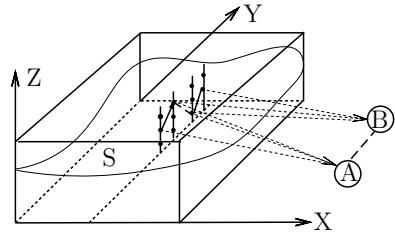


Fig. 2. Surface S and nodes A and B

Stitching of CuBOSS. To form an optimal surface from S constructed in the above section, our idea is to iteratively improve S by local adjustment called stitching. In this section, we first interpret what is stitching and then show how to iteratively apply it to obtain an optimal surface. Firstly, it is easy to see

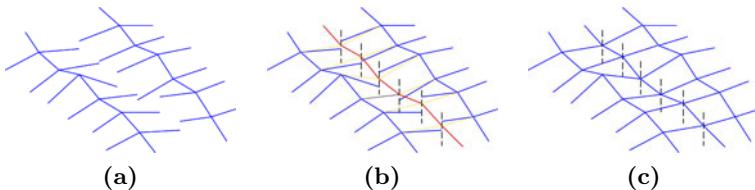


Fig. 3. (a) Before stitching (b) Stitching (C) After stitching

that before S achieves optimum, there must exist at least one x-slice $S(i, :, :)$ on which the voxels are not entirely located at optimal positions. Otherwise S must be optimal. Thus we follow the same idea in [9] to adjust $S(i, :, :)$ locally when we fix the rest of S other than $S(i, :, :)$. However, the surface could not be detected if the maximum allowed curvature of the surface is small. To overcome this difficulty, instead of considering a small subset of voxels on each column of $S(i, :, :)$, all voxels on the column are converted to the nodes in $G_s(i)$ in the following way. Let $S_l(1 : i - 1, :, :)$ and $S_r(i + 1 : n_x, :, :)$ be two partial surfaces. To stitch $S_l(1 : i - 1, :, :)$ and $S_r(i + 1 : n_x, :, :)$ in slice $I(i, :, :)$ (see Figure. 2 (a)-(c)), we fix the selected voxels in slices $I(i - 1, :, :)$ and $I(i + 1, :, :)$, and convert

the stitching problem into a 2D shortest path problem in graph $G_s(i)$ which is similar to the graph G_1 discussed in the above section. $G_s(i)$ has the same size as G_1 but with different cost functions including the weight and the mean curvature of voxels in slice $I(i, :, :)$, as well as the adjustment of the x-curvatures for those selected voxels in slices $I(i-1, :, :)$ and $I(i+1, :, :)$. We replace equation (1) by (2). (Two terms of (2) are represented in (3) and (4) respectively.)

$$\text{cost}[V(u, c, b)] = \alpha \text{costw}[V(u, c, b)] + (1 - \alpha) \text{costk}[V(u, c, b)], \quad (2)$$

$$\begin{aligned} \text{costw}[V(u, c, b)] = & w(x_c, y_c, z_c) + w(i-1, y_c, z_l) + w(i+1, y_c, z_r) \\ & + 2(w(x_c, y_c, z_c) - w(x'_c, y'_c, z'_c)) \end{aligned} \quad (3)$$

$$\text{costk}[V(u, c, b)] = \kappa[V(u, c, b)] + \Delta\kappa[I(i-1, y_c, z_l)] + \Delta\kappa[I(i+1, y_c, z_r)], \quad (4)$$

$I(i-1, y_c, z_l)$ (or $I(i+1, y_c, z_r)$) is the voxel of S on column $\text{Col}(i-1, y_c)$ (or $\text{Col}(i+1, y_c)$). $w(x'_c, y'_c, z'_c)$ is the weight cost of voxel on column $\text{Col}(i, y_c)$ of unadjusted $S(i, :, :)$. We set coefficient 2 before $(w(x_c, y_c, z_c) - w(x'_c, y'_c, z'_c))$ in equation (3) to reflect the weight adjustment for both selected voxels in slices $I(i-1, :, :)$ and $I(i+1, :, :)$. In equation (4), we include the curvature of voxels in slice $I(i, :, :)$ as well as adjustment of the x-curvatures for both selected voxels in slices $I(i-1, :, :)$ and $I(i+1, :, :)$. Thus an locally improved $S(i, :, :)$ can be obtained by shortest path algorithm on graph $G_s(i)$ within the same running time $O(n_y n_z N^2)$ on G_1 .

Since stitching on one x-slice could locally improve S , our strategy is repeatedly apply the stitching algorithm on the obtained surface with each time shifting the to-be-stitched slices to the surface with a significant improvement in the previous iteration. By setting a threshold τ close to 0, we then (1) detect the non-stable portions of the surface, (2) apply the stitching on both x and y-slices, (3) compare the two obtained surfaces and (4) use the better one as the solution. Compared to the method in [9], in which if the current iteration of stitching is on slices $I(1, :, :), I(3, :, :), \dots, I(2i-1, :, :), \dots$, then in the next iteration, the stitching is on slices $I(2, :, :), I(4, :, :), \dots, I(2i, :, :), \dots$, our approach has the similar property as in [9] (i.e., the cost of the surface will monotonically decrease in each iteration) but with less running time. In all our experiments with the OCT images, our approach converges to the optimal (or almost optimal) within a small constant (3-5) number of iterations.

4 Experimental Results

To assess the empirical performance of the algorithm, we focus on two types of data: (a) a set of computer generated phantom images with controlled noise level (b) clinical data (where results of our method are compared with manual segmentation provided by the experts). We demonstrate the expected behavior of our method with respect to its performance as a function of different parameters

(e.g. N and α). In all the experiments, we set τ as 5 which is less than 1% of the minimum surface value of one slice among different α s .

Synthetic Datasets. For synthetic data, we used computer to generate 12 sets of phantoms images in the following way. The size of the phantom images mimicked the size of real images, i.e. $(100 \times 100 \times 128)$. Single terrain-like surface was embedded in a volumetric image. Voxels belonging to the surface have higher gray values in contrast with the white background. To make the problem more realistic phantom datasets were blurred and superimposed with a Gaussian noise, for $\sigma_{noise} \in \{0.1, 1\}$. The table below summarizes the unsigned and signed errors in voxel size.

σ_{noise}	Unsigned Positioning Error		Signed Positioning Error	
	μ_u	σ_u	μ_s	σ_s
0.1	0.60	0.26	0.02	0.54
1	0.79	0.34	0.09	0.69

OCT Images. The proposed algorithm was examined on 14 datasets of 3D OCT images $(200 \times 200 \times 256)$ with the voxel size $(6 \times 6 \times 2\mu m)$. The average of the two tracings from two human experts were used as the reference standard. The unsigned surface positioning errors were computed for each surface as a sum of absolute values of distance between the computed results and the reference standard. (2.4 GHz, 4GB memory).

In our experiments, we apply our algorithm to the down-sampled 3D images $(100 \times 100 \times 128)$. The results were up-sampled to the size of original 3D images before we compare them to the reference standard. Our algorithm is implemented by C++ and LEDA (Library of Efficient Data types and Algorithms)- 5.2. All of the experiments were conducted on a Linux workstation. By iterative application of our method we could obtain results for several surfaces. Here, we analyze in details results for surface 1 which was the first surface detected by our algorithm. The computed unsigned position errors in μm are summarized in the table to the right where $N = 3$.

α	Unsigned Positioning Error	
	μ	σ
0.25	4.84	2.94
0.50	4.88	3.01
0.75	4.20	2.68
1.00	4.96	2.64

Our segmentation results are comparable to previous results reported in the literature [9,4,5]. In Figure 5, three cross-sections that show overlay of the results of our method (in green) and contours provided by the experts (in red). In Figure 4(a), the average running time is increasing with respect to N . Note that even for a rather high value of the parameter N (e.g. 5) the average running time remains around 4 minutes. It is much less than the reported running time (i.e., 10 or 20 minutes). To explore the role of α , Figure 4(b) (or (c)) shows the change of

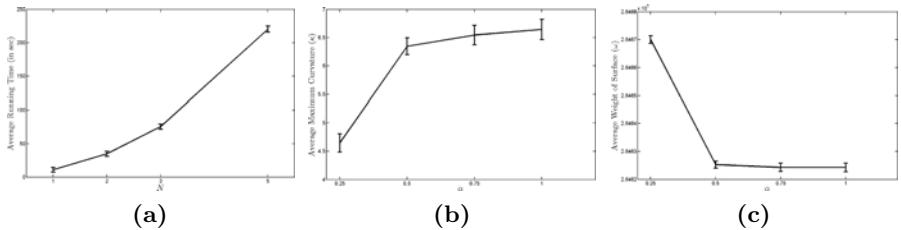


Fig. 4. (a) Average running time with respect to smoothness parameter N . (b) Average maximum curvature as a function of α . (c) Average weight of the surface as a function of α .

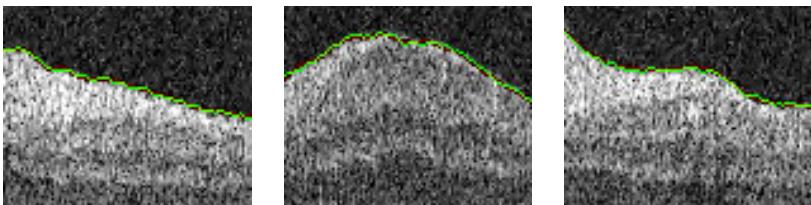


Fig. 5. Three different slices containing comparison of our algorithm results (in green) with the manual segmentation (in red)

maximum surface curvature (or the surface weight) as a function of α . We can see that the maximum curvature of the surface increases and the total surface weight decreases with the increase of α .

5 Discussion and Conclusion

We have presented a novel algorithm for segmenting globally optimal and smooth surface. Comparing with the current best approaches [9,5], our method has comparable segmentation results with a much improved running time, $O(n_y n_z N^2)$ theoretically and several minutes practically. The correct surface may not be detected if the maximum curvature allowed in the surface is small in [9]. Our approach successfully overcomes this difficulty since there are no restrictions on the curvature upper bound. The proposed method is suitable for fully parallel implementation on GPUs, which could potentially allow us to segment highly noisy volumetric images in real time.

References

1. Ardon, R., Cohen, L., Yezzi, A.: A new implicit method for surface segmentation by minimal paths: Applications in 3D medical images. In: Rangarajan, A., Vemuri, B.C., Yuille, A.L. (eds.) EMMCVPR 2005. LNCS, vol. 3757, pp. 520–535. Springer, Heidelberg (2005)

2. Boykov, Y., Kolmogorov, V.: Computing geodesics and minimal surfaces via graph cuts. In: ICCV, pp. 26–33 (2003)
3. Felzenszwalb, P., Huttenlocher, D.: Efficient graph-based image segmentation. Int. J. Comput. Vision 59(2), 167–181 (2004)
4. Garvin, M., Abramoff, M., Kardon, R., Russell, S., Wu, X., Sonka, M.: Intraretinal Layer Segmentation of Macular Optical Coherence Tomography Images Using Optimal 3-D Graph Search. IEEE Trans. Med. Imaging 27(10), 1495–1505 (2008)
5. Haeker, M., Wu, X., Abràmoff, M., Kardon, R., Sonka, M.: Incorporation of regional information in optimal 3-D graph search with application for intraretinal layer segmentation of optical coherence tomography images. In: Karssemeijer, N., Lelieveldt, B. (eds.) IPMI 2007. LNCS, vol. 4584, pp. 607–618. Springer, Heidelberg (2007)
6. Leventon, M., Grimson, W., Faugeras, O., Wells, W.: Level Set Based Segmentation with Intensity and Curvature Priors. In: IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (2000)
7. Wyatt, C., Ge, Y., Vining, D.: Segmentation in virtual colonoscopy using a geometric deformable model. In: Computerized Medical Imaging and Graphics, vol. 30(1), pp. 17–30 (2006)
8. Wu, X., Chen, D.Z.: Optimal net surface problems with applications. In: Widmayer, P., Triquero, F., Morales, R., Hennessy, M., Eidenbenz, S., Conejo, R. (eds.) ICALP 2002. LNCS, vol. 2380, pp. 1029–1042. Springer, Heidelberg (2002)
9. Xu, L., Stojkovic, B., Zhu, Y., Song, Q., Wu, X., Sonka, M., Xu, J.: Efficient algorithms for segmenting globally optimal and smooth multi-surfaces. In: IPMI (2011)

Automated Nuclear Segmentation of Coherent Anti-Stokes Raman Scattering Microscopy Images by Coupling Superpixel Context Information with Artificial Neural Networks

Ahmad A. Hammoudi^{1,2}, Fuhai Li¹, Liang Gao^{1,3}, Zhiyong Wang¹, Michael J. Thrall⁴, Yehia Massoud², and Stephen T.C. Wong^{1,4,*}

¹ Department of Systems Medicine and Bioengineering, The Methodist Hospital Research Institute, Weill Cornell Medical College, Houston, TX
`{aahammoudi, fli, lgao2, stwong}@tmhs.org`

² Department of Electrical and Computer Engineering, Rice University, Houston, TX

³ Department of Bioengineering, Rice University, Houston, TX

⁴ Department of Pathology and Laboratory Medicine, The Methodist Hospital and Weill Cornell Medical College, Houston, TX

Abstract. Coherent anti-Stokes Raman scattering (CARS) microscopy is attracting major scientific attention because its high-resolution, label-free properties have great potential for real time cancer diagnosis during an image-guided-therapy process. In this study, we develop a nuclear segmentation technique which is essential for the automated analysis of CARS images in differential diagnosis of lung cancer subtypes. Thus far, no existing automated approaches could effectively segment CARS images due to their low signal-to-noise ratio (SNR) and uneven background. Naturally, manual delineation of cellular structures is time-consuming, subject to individual bias, and restricts the ability to process large datasets. Herein we propose a fully automated nuclear segmentation strategy by coupling superpixel context information and an artificial neural network (ANN), which is, to the best of our knowledge, the first automated nuclear segmentation approach for CARS images. The superpixel technique for local clustering divides an image into small patches by integrating the local intensity and position information. It can accurately separate nuclear pixels even when they possess subtly lower contrast with the background. The resulting patches either correspond to cell nuclei or background. To separate cell nuclei patches from background ones, we introduce the rayburst shape descriptors, and define a superpixel context index that combines information from a given superpixel and its immediate neighbors, some of which are background superpixels with higher intensity. Finally we train an ANN to identify the nuclear superpixels from those corresponding to background. Experimental validation on three subtypes of lung cancers demonstrates that the proposed approach is fast, stable, and accurate for segmentation of CARS images, the first step in the clinical use of CARS for differential cancer analysis.

Keywords: Coherent anti-Stokes Raman scattering (CARS) microscopy, Nuclear segmentation, Superpixels, Artificial Neural Network (ANN).

* Corresponding author.

1 Introduction

Lung cancer is the primary cause of cancer deaths in the United States with 222,500 new cases of lung cancer and 157,300 lung cancer deaths in 2010 [1]. The five-year survival rate is less than 18% [2, 3]. Although early detection has attracted major research interest [4, 5], less than 1% of patients can be diagnosed at an early stage [6]. Tissue biopsy is frequently needed as a follow-up test upon pulmonary examination using computed tomography (CT) and magnetic resonance imaging (MRI) for definitive diagnosis. However, it remains difficult to obtain samples precisely at the site of small lesions [6]. Some patients will need to undergo re-biopsy, resulting in increased costs and delays in diagnosis and treatment. Given the risks and cost of lung biopsy, it would be beneficial to develop an imaging strategy that can provide real time images together with diagnostic analysis of the biopsied site and thus relieve the difficulty in finding small lesions, limit damage to lung tissue, diagnose lung cancer *in vivo*, and provide diagnostic yield comparable to existing biopsy methods.

The Coherent anti-Stokes Raman scattering (CARS) imaging technique [7] holds great promise for this diagnostic application. It captures intrinsic molecular vibrations to create optical contrast with submicron level spatial resolution, as well as video-speed imaging rate [8]. In the CARS process, a pump field (ω_p), a Stokes field (ω_s) and a probe field (ω_p') interact with the samples through a four-wave mixing process [9]. When the frequency difference, $\omega_p - \omega_s$ (beating frequency), is in resonance with a molecular eigenvibration, an enhanced signal at the anti-Stokes frequency, $\omega_{\alpha\sigma} = \omega_\pi - \omega_\sigma + \omega_p'$, is generated [10]. The major advantage of CARS is that the signal yield is much higher, typically several orders of magnitude, than the signal yield obtained through the conventional spontaneous Raman scattering process [11]. As a result, this imaging technique has been used to visualize different tissue structures, e.g. skin [11], lung and kidney [8].

In this study, we investigate the use of CARS for lung cancer imaging, and perform the necessary groundwork required to exploit the potential of CARS for differential diagnosis of lung cancer subtypes. Figure 1 shows representative images of three subtypes of lung cancers (adenocarcinoma, small cell carcinoma and squamous cell carcinoma), these three cancer subtypes possess distinct pathologic features with regards to cell size, density and orientation in CARS images. Thus, nuclei segmentation will be a critical step to extract this information, crucial for the differential diagnostic analysis. Moving forward in conjugating CARS with existing image-guided-biopsy setups necessitates an automatic quantitative image analysis system, capable of performing the aforementioned segmentation and information extraction task, to enable real time interpretation of the acquired CARS images. As opposed to manual segmentation, which is time-consuming and subjective, an automated nuclear segmentation approach is highly attractive and will bring significant contributions to the translation of the CARS technology in to clinical diagnosis.

However, the low signal-to-noise ratio (SNR) and uneven background of CARS images have prevented effective automated segmentation using existing nuclear segmentation approaches, such as simple thresholding [3], voronoi tessellation [4], seeded watershed [5], graph cut [6], and active contours [7]. None of them could

effectively perform nuclear segmentation on CARS images without significant adjustments. For example, simple thresholding or adaptive thresholding cannot separate cell pixels with low intensity contrast from the background, and cannot separate touching cells. The use of marker controlled watershed, graph cuts, and active contour approaches all require the presence of detected cell centers, which is another challenging problem for CARS images. In addition, the Voronoi method can only generate the rough regions of cell nuclei. Marker controlled watershed delineates nuclear boundaries on the pixels with maximum intensity between two nuclei. Thus it is biased to the intensity variation. Graph cut methods partition images through regions of sharp variations, e.g. Edges or an intensity change. However, the intensity variation and uneven background can mislead the graph cuts. Active contours require good initial contours that are close to the real boundaries, otherwise the intensity variation will result in edge leaking and early edge stop.

To circumvent the aforementioned challenges, in this paper, we propose a novel fully automated nuclear segmentation method for CARS images. The flowchart of the proposed method is shown in Fig.2. The key idea of the proposed method is to delineate cell boundaries using local clustering, and over segment the image into superpixels, this overcomes the low SNR and uneven background. Then, separate the clusters (superpixels) corresponding to cell nuclei from those corresponding to background superpixels, this is done by introducing a superpixel context index and rayburst shape descriptors as the inputs of ANN classifier. The rest of the paper is organized as follows. The details of the proposed method are presented in Section 2. Section 3 provides the experimental validation. Conclusion and discussion are in Section 4.

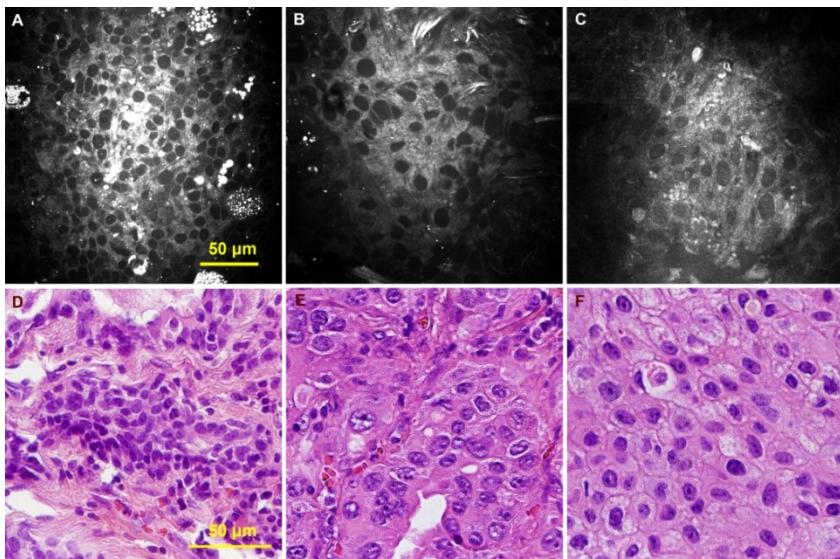


Fig. 1. Representative CARS images (upper panels) and corresponding H&E images (lower panels) images of three lung cancer subtypes: small cell carcinoma (A and D), adenocarcinoma (B and E), and squamous cell carcinoma (C and F)

2 Methodology

A. Local clustering. It has been shown that using both local intensity and position information of image patches could greatly enhance cell identification and segmentation in microscopic images, especially in images with large numbers of closely positioned cells, low SNR, and uneven background [12]. Superpixels have been shown in [13] and [14] to have superior performance to using rectangular image patches, for localized image processing. Therefore in this work, we employ Simple Linear Iterative Clustering (SLIC) [13] to divide the CARS images into small patches (superpixels). Before performing local clustering, images were preprocessed with adaptive histogram equalization and a Gaussian filter to enhance performance by reducing noise and making the intensity more uniform. The clustering algorithm makes use of local intensity information, and introduces a distance measure to control the compactness of each cluster, which produces clusters that adhere well to the image boundaries, and thus the boundaries of the nuclei. Figures 3, 4, 5 show the superpixels results obtained by performing SLIC on three subtypes of lung cancer. The resulting image superpixels show a good segmentation of an image into regions corresponding to intact nuclei and background. Consequently, the nuclear segmentation task becomes a pattern recognition problem – to discriminate nuclear superpixels from background regions.

B. Superpixel Context index and Rayburst Shape Descriptors. A set of informative quantitative features are needed for the identification of nuclear superpixels. Due to the uneven background intensity and low SNR, it is difficult to discriminate nuclei from background only using the intensity information of individual superpixels. Since the nuclear superpixels have, in general, regular shapes, and low and uniform intensity distribution compared to the background superpixels, the nuclei within a small neighborhood are distinguishable mainly through two major factors: The intensity variation of a superpixel with respect to its neighbors, and the shape uniformity. Thus we define the superpixel context index and rayburst shape descriptor for distinguishing the nuclei from background.

We calculate two superpixel context indices as follows. For each superpixel, locate all of its immediate neighbor superpixels, and then calculate the ratios of the average and median intensity of the superpixel to the average and median intensities of its neighboring superpixels respectively. We set the superpixel context indices as the minimum ratio values (mean and median) based on the fact that at least one immediate neighbor is a background superpixel.

To measure the shape uniformity, we design the rayburst shape descriptors as follows: locate the centroid of a given superpixel, then shed ray lines to the boundary points uniformly (each at a 10° arc), then we use the standard deviation of lengths of the 36 ray lines as the uniformity value.

In addition, we add two other relevant descriptors that can demonstrate shape uniformity. The first one is the goodness of fit between the superpixel and the best fitting ellipse. We fit each superpixel with an ellipse, and then calculate the relative regions

outside the ellipse and missing regions inside the ellipse compared to the ellipse size. The second one is the ratio of superpixel area to the area of its convex hull. We end up with five features for distinguishing the nuclear superpixels from background.

C. Nuclear Superpixel Identification Using ANN. After selecting a set of features, another important step is to select a good classifier. In this study, we choose an ANN classifier, which could learn the optimal feature combinations for the nuclear superpixel identification. In this study, a neural network with 100 neurons and one hidden layer is trained for each of the lung cancer subtypes. Figure 2 provides the complete segmentation and detection procedure, the feature vector elements f^1 thru f^5 are the five aforementioned distinguishing features. To train the network, we manually labeled the superpixels (as nucleus or background) in three images for each subtype of lung cancer to train the three ANN classifiers respectively.

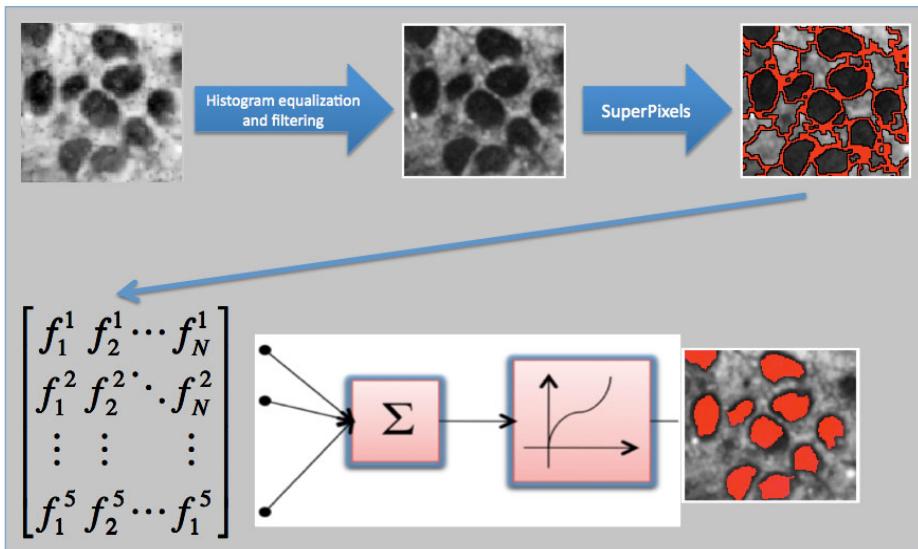


Fig. 2. Flowchart of the proposed segmentation approach. It starts with filtering and adaptive histogram equalization, generates superpixels, creates features and performs classification.

3 Experimental Validation

To evaluate the proposed approach, we randomly selected images for each lung cancer subtype, to which we applied the proposed segmentation approach. Figures 3, 4, 5 provide the representative segmentation sample results of adenocarcinoma, small cell

carcinoma and squamous cell carcinoma subtypes respectively. To validate the segmentation accuracy, we compared the automated segmentation results with manual segmentation results (the ground truth) by calculating three scores: precision, recall and fscore, as follows:

$$p = \frac{S_i \cap S'_i}{S'_i}, \quad r = \frac{S_i \cap S'_i}{S_i}, \quad f = \frac{2 \times p \times r}{p + r}$$

Where S_i is the ground truth of manually labeled cells and S'_i is the automated segmentation result. Figure 6 shows validation of the segmentation results in terms of precision, recall and fscore. As we can see that all three indices are above 90%, which indicates the proposed nuclear segmentation method of CARS images is stable and accurate for all the three subtypes of lung cancer.

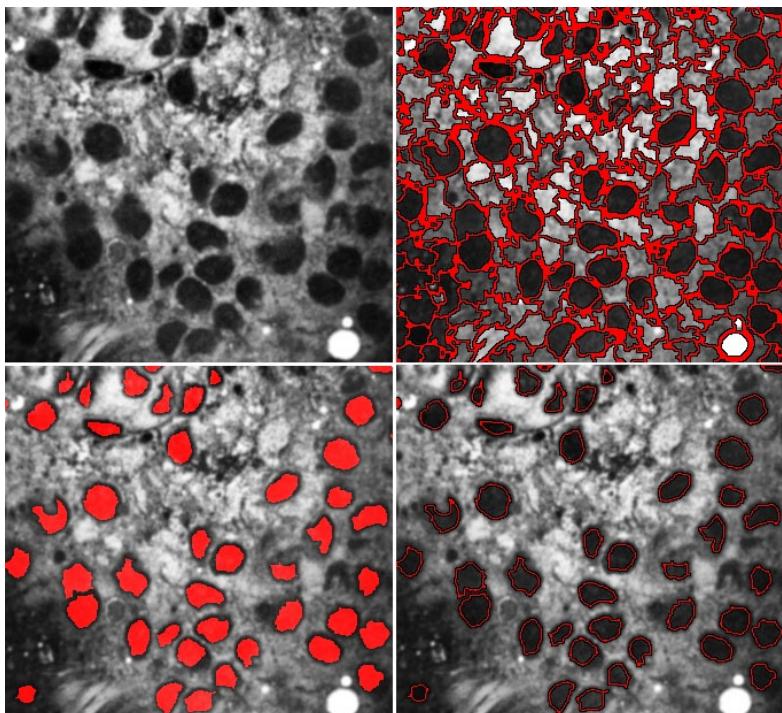


Fig. 3. Representative segmentation results of Adenocarcinoma subtype of lung cancer. Showing original image (top left), superpixels (top right), and segmentation (bottom). The red masks and curves indicate cells and their boundaries.

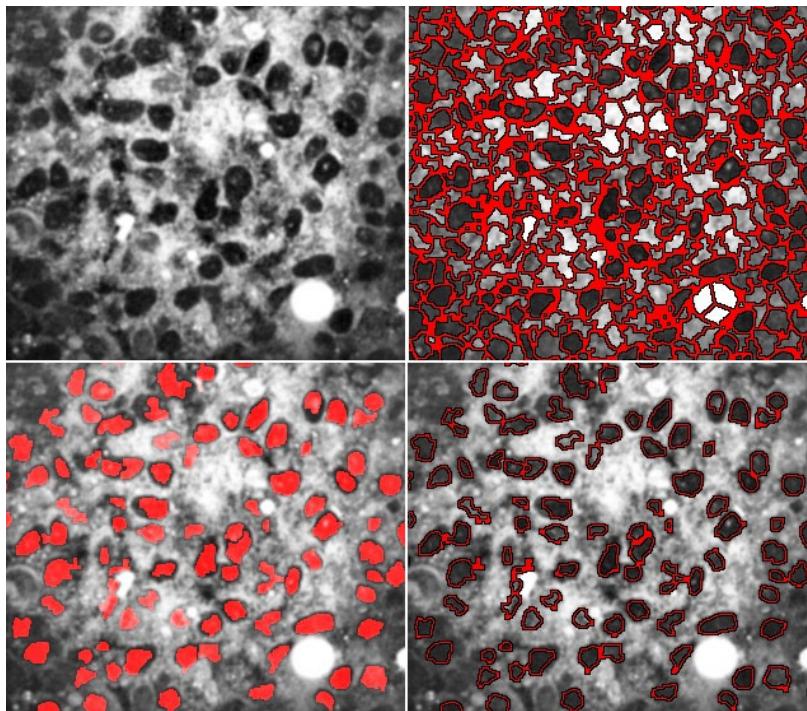


Fig. 4. Representative segmentation results of Small cell carcinoma subtype of lung cancer. Showing original image (top left), superpixels (top right), and segmentation (bottom). The red masks and curves indicate cells and their boundaries.

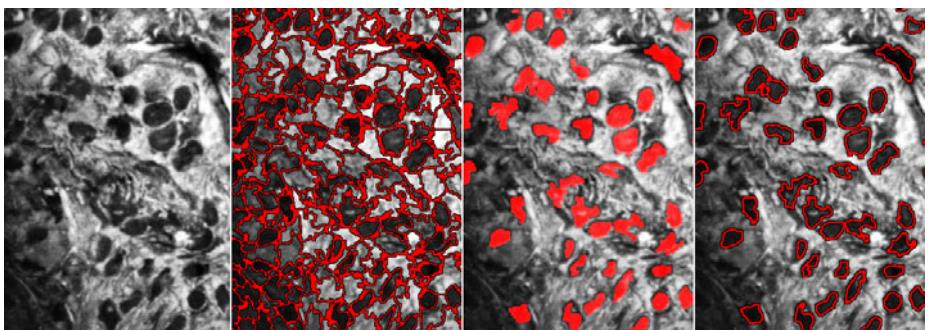


Fig. 5. Representative segmentation results of Squamous cell carcinoma subtype of lung cancer. Showing original image (far left), superpixels (second from left), and segmentation (second and far right). The red masks and curves indicate cells and their boundaries.

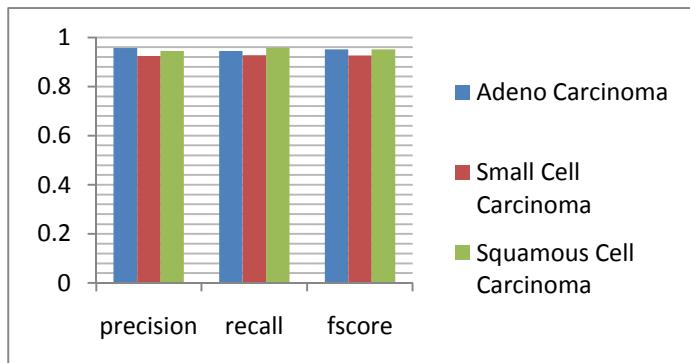


Fig. 6. Segmentation accuracy of the proposed method on three subtypes of lung cancer

4 Conclusion and Future Work

CARS is a novel imaging technique, with the potential for clinical use in differential diagnosis of cancer, for which automatic segmentation of nuclei is a key component of developing diagnostic algorithms. However, existing segmentation approaches cannot effectively overcome the low SNR and uneven background challenges in CARS images for nuclear segmentation. In this paper, we propose a fully automated nuclear segmentation method, coupling superpixels (local clustering) and ANNs, for identification of cell nuclei in CARS lung cancers images. Thus, we provide a solution to a critical step for differential analysis using cell morphology. The experimental results show that the proposed automated method possesses great capability for nuclear segmentation on CARS images. To further extend our work, we intend to enhance the selection of identification features to make sure we can detect irregularly shaped nuclei, and solve some remaining over segmentation and under segmentation issues. To get to the full potential of using images for differential analysis, the local clustering and classification technique is to be extended to 3D cars images, where biologically significant features can be extracted to classify segmented cells into their respective cancer subtypes.

Acknowledgment. This work was completed with grant support from the John S Dunn Research Foundation.

References

1. Hashizume, H., Baluk, P., Morikawa, S., McLean, J.W., Thurston, G., Roberge, S., Jain, R.K., McDonald, D.M.: Openings between defective endothelial cells explain tumor vessel leakiness. *Am. J. Pathol.* 156, 1363–1380 (2000)
2. Youlden, D.R., Cramb, S.M., Baade, P.D.: The International Epidemiology of Lung Cancer: geographical distribution and secular trends. *J. Thorac. Oncol.* 3, 819–831 (2008)
3. Parkin, D.M., Bray, F., Ferlay, J., Pisani, P.: Global cancer statistics. *CA Cancer J. Clin.* 55, 74–108 (2005)

4. Diederich, S.: Lung cancer screening: status in 2007. *Radiologe* 48, 39–44 (2008)
5. Henschke, C.I., Yankelevitz, D.F., Libby, D.M., Pasmanier, M.W., Smith, J.P., Miettinen, O.S.: Survival of patients with stage I lung cancer detected on CT screening. *N. Engl. J. Med.* 355, 1763–1771 (2006)
6. McWilliams, A., MacAulay, C., Gazdar, A.F., Lam, S.: Innovative molecular and imaging approaches for the detection of lung cancer and its precursor lesions. *Oncogene* 21, 6949–6959 (2002)
7. Duncan, M.D., Reintjes, J., Manuccia, T.J.: Scanning coherent anti-Stokes Raman microscope. *Optics Letters* 7, 350–352 (1982)
8. Evans, C.L., Xie, X.S.: Coherent Anti-Stokes Raman Scattering Microscopy: Chemical Imaging for Biology and Medicine. *Annu. Rev. Anal. Chem.* 1, 883–909 (2008)
9. Cheng, J.-X., Xie, X.S.: Coherent Anti-Stokes Raman Scattering Microscopy: Instrumentation, Theory, and Applications. *J. Phys. Chem. B* 108, 827–840 (2004)
10. Evans, C.L., Potma, E.O., Xie, X.S.: Coherent anti-stokes raman scattering spectral interferometry: determination of the real and imaginary components of nonlinear susceptibility $\chi(3)$ for vibrational microscopy. *Optics Letters* 29, 2923–2925 (2004)
11. Evans, C.L., Potma, E.O., Puoris'haag, M., Cote, D., Lin, C.P., Xie, X.S.: Chemical imaging of tissue in vivo with video-rate coherent anti-Stokes Raman scattering microscopy. *Proc. Natl. Acad. Sci. USA* 102, 16807–16812 (2005)
12. Zhaozheng, Y., Bise, R., Mei, C., Kanade, T.: Cell segmentation in microscopy imagery using a bag of local Bayesian classifiers. In: 2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro, pp. 125–128 (2010)
13. Radhakrishna, A., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: SLIC Superpixels. Technical Report 149300, EPFL (June 2010)
14. Lucchi, A., Smith, K., Achanta, R., Lepetit, V., Fua, P.: A fully automated approach to segmentation of irregularly shaped cellular structures in EM images. In: Jiang, T., Navab, N., Pluim, J.P.W., Viergever, M.A. (eds.) MICCAI 2010. LNCS, vol. 6362, pp. 463–471. Springer, Heidelberg (2010)

3D Segmentation in CT Imagery with Conditional Random Fields and Histograms of Oriented Gradients

Chetan Bhole¹, Nicholas Morsillo¹, and Christopher Pal²

¹ University of Rochester

² École Polytechnique de Montréal

Abstract. In this paper we focus on the problem of 3D segmentation in volumetric computed tomography imagery to identify organs in the abdomen. We propose and evaluate different models and modeling strategies for 3D segmentation based on traditional Markov Random Fields (MRFs) and their discriminative counterparts known as Conditional Random Fields (CRFs). We also evaluate the utility of features based on histograms of oriented gradients or HOG features. CRFs and HOG features have independently produced state of the art performance in many other problem domains and we believe our work is the first to combine them and use them for medical image segmentation. We construct 3D lattice MRFs and CRFs, use variational message passing (VMP) for learning and max-product (MP) inference for prediction in the models. These inference and learning approaches allow us to learn pairwise terms in random fields that are non-submodular and are thus very flexible. We focus our experiments on abdominal organ and region segmentation, but our general approach should be useful in other settings. We evaluate our approach on a larger set of anatomical structures found within a publicly available liver database and we provide these labels for the dataset to the community for future research.

Keywords: MRF, CRF, generative, discriminative, 3D segmentation, HOG.

1 Introduction

Accurate detection and segmentation of organs, vessels and other regions of interest is a key problem in medical imaging. Markov random fields (MRFs) provide an attractive framework for image segmentation. Recent insights into modeling techniques have given rise to a new distinction between traditional MRFs which define a joint distribution over both segmentation classes and features and conditional random fields (CRFs) which model the conditional distributions of the segmentation field directly. CRFs have had a major impact in machine learning in recent years and our work here explores their application to 3D image segmentation in detail. Image descriptors based on histograms of oriented image gradients or HOG features have received a lot of attention in computer vision recently. For example, both the widely used SIFT descriptors of Lowe [10] and the person detector of Dalal et al. [4] are based on different forms of HOG. Depending on the size and spatial extent of the HOG, such features are capable of capturing the entire shape of small organs or contour segments of larger organs. In our work here, we propose and explore a HOG based feature descriptor specifically designed for detecting

anatomical structures and contours in CT imagery of the abdomen. To visualize some key elements of our approach, we show an example of a preprocessed axial image slice, a manual segmentation into organs of interest and some pre-processing steps in Fig. 1. We model image volumes using the 3D graphical models as shown in Fig. 2. Each 2D lattice corresponds to an axial slice of the medical image.

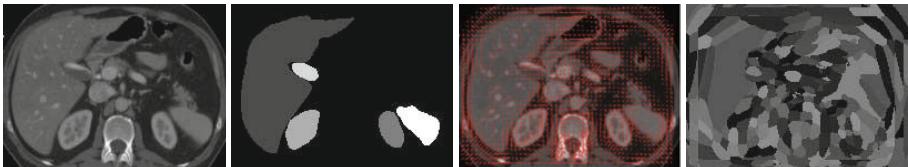


Fig. 1. Left to right: A coarsely registered axial data slice and its manual segmentation. Each shade denotes a different class label. The third image is a visualization of the HOG features (zoom for clarity). Arrows indicate gradient orientations and magnitude. Only prominent gradient bins are shown to reduce clutter. The image on the right is the classification using HOG codewords.

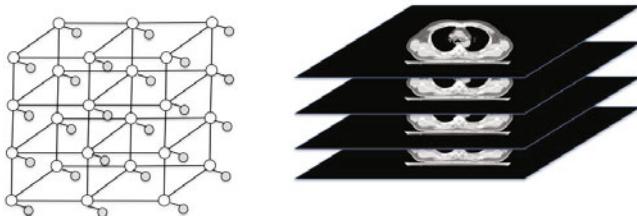


Fig. 2. The figure shows a graphical model modeling the image volume. The shaded nodes denote the observed feature node variables associated with a pixel of an image. The unshaded nodes are the segmentation variables representing the anatomy classes.

Markov Random Fields have a long history in image processing and computer vision; however, they started to receive more intense attention in the medical image analysis around the time of Zhang, Brady and Smith's influential work [16] on hidden Markov Random Fields for segmenting Magnetic Resonance (MR) imagery of the brain. They created a traditional MRF in the sense that in that they created a likelihood term for each pixel and a spatially coupled prior in the form of a random field. One of their contributions was to propose and explore the use of Gaussian mixture models for the local likelihood terms, thus introducing hidden variables and producing a hidden MRF. They then used the Expectation Maximization (EM) algorithm to perform unsupervised learning in the hidden MRF for segmentation. While exploring the problem of general interactive image segmentation Blake et al. [1] used a similar Gaussian mixture model based random field (GMMRF) approach along with a small amount of training data. This allowed one to learn MRF models in a supervised manner; however, the model still had the form of a traditional MRF defining the joint probability of features and segmentation class labels. They however do not learn pairwise term parameters. Lafferty, McCallum and Pereira's landmark work [7] on conditional random fields (CRFs)

presented a new MRF approach based on directly modeling the conditional distribution of a field of labels. Their work focused on the use of chain structured CRFs and applications in natural language processing. Soon after Kumar and Hebert [6] explored the use of two dimensional lattice structures for general image segmentation.

While there has been an explosion of activity in the general computer vision community using CRFs, there has been comparatively less exploration of CRFs in the medical image segmentation literature. Some notable previous work includes Tsechpenakis et al. [13] coupled CRFs with deformable models for 3D eye segmentation. The deformable model captures shape information and is fed to the CRF model as observations. Lee et al. [8] used pseudo-CRFs for segmenting brain tumors. As global inference in a true CRF can be expensive during learning, they broke the problem up into two components, one of them not depending on spatial interactions and another term that accounted for spatial interactions which they view as a regularizer.

While modeling choices are a key part of solving the problem of image segmentation, the choice of image features is equally critical. Image features based on HOG features [4] have generated considerable interest in computer vision for tasks such as human detection in photographs. While HOG techniques have received an explosion of exploration in computer vision there has been comparatively little work in medical image analysis. There has been some recent work using HOG like techniques such as [11] which used a HOG like feature in their work on segmenting brain structures from Diffusion Tensor (DT-MR) images. Graf et al. [5] recently explored the use of a pyramidal HOG for detecting vertebrae in 2D CT imagery. We are motivated to use HOG based features here as oriented gradients provide a way to capture shapes and partial curves through gradient profiles which can be more robust to variations across patients and different spatial contexts. To the best of our knowledge, we provide the first exploration of HOG features for multi-organ 3D CT image segmentation. Our experimental evaluation confirm their utility in our setting here.

Other exemplary work on organ segmentation such as Siebert et al. [12] used landmark detectors consisting of 3D Haar features for image volumes. They target full body segmentation and use 6 organs for segmentation. The landmarks are used in an MRF framework to obtain organ centers. After obtaining organ centers, marginal space learning classifiers which are a sequence of Probabilistic Boosting Trees are used to perform organ segmentation. They achieve full body organ localization and segmentation using a hierarchical and contextual approach. In contrast, Ling et al. [9] used a hierarchical approach, marginal space learning and steerable features to segment a liver and improve upon previous methods of shape initializations. Other recent work [3] has used regression forests to detect and localize abdominal organs. Regression forests could be integrated into the MRF approaches we explore here. Varshney et al. [14] provides a survey of the different approaches that have been used for segmentation of abdominal organs. Neural networks, level set methods, model fitting and rule based methods have received particular attention.

We make a number of contributions in this paper. First, we demonstrate how discriminative parameter learning in CRFs indeed leads to better segmentation performance compared to than their generative MRF counterparts. Second, we demonstrate how HOG features are indeed well suited to the anatomical segmentation problem.

Third, we model pairwise terms using non-submodular functions and perform full 3D inference using VMP and MP.¹

2 Our Approach

We use 22 3D volumes of patients to perform our experiments. 19 of these volumes are taken from the liver data set (<http://www.sliver07.org/>) and we provide the additional labels for the dataset to the community for future research. We use the cases that had segmented liver results provided. Our goal is to segment five organs from the background: the liver, two kidneys, spleen and gall bladder. In the beginning, we very coarsely align the volumes using 6 manually selected landmark key points and a global 3D affine transformation. This is done because there is a lot of variation in the images regarding the relative positions of these organs. 2 points capture the axial extent (height) of the abdominal region of interest. We then select 4 landmark points on an image slice that has the first appearance of the right kidney as we look at the slices from top to bottom. Two of these points are marked on the two sides (left and right) of the sternum and the other 2 points are marked on the top of the rib cage and bottom of the vertebra. We use linear interpolation so that the resultant scan has 150 slices and an image size of 300(width) x 400(height). We scale the image set down by a half on all dimensions in our experiments for computational reasons. We use a higher range of intensity values (-180 to 1200HU) so as to be able to accommodate any contrast enhancement or gall bladder or kidney stones. We quantize intensity values into bins and model intensity as a histogram or discrete distribution. The gradient between two neighboring pixels is also modeled as a histogram of gradient values. The gradient is the difference of pixel intensities of the neighboring pixels. We model the spatial location of organs using a mixture of Gaussians and the choice of the number of Gaussians in the mixture is determined by cross validation set likelihood and we note that the results tend to depend on the organ's size and compactness. Appearance features are modeled as a 5x5 patch around each pixel and clustering is used. HOG features are also clustered to create code books. To compute the HOG features we use non overlapping blocks where each block is made of 3x3 cells and each cell of 8x8 pixels has 10 orientation bins.

The feature variables for the pixel p are x_p^{int} , x_p^{app} , x_p^{hog} and x_p^{loc} to denote features of intensity, appearance, HOG and location respectively. The gradient feature variable for neighboring pixels p and q is given by x_{pq}^{grad} . We depict a discriminative component with an edge and a dark node on the edge and a generative component with a directed edge so that the arrow points to the variable that is generated.

We look at two different probabilistic models for image segmentation. The models are either fully discriminative CRFs or contain both generative and discriminative components (CRF-MRF). Fig 3 shows the model structures.

A conditional random field [7] can be expressed as an undirected graph or random field which has associated with it a conditional distribution $p(\mathbf{y}|\mathbf{x})$ where \mathbf{y} denotes the output variables (e.g. image segmentation classes) and \mathbf{x} denotes the input variables (such as features of an image). Let $\mathbf{x} \equiv [\mathbf{x}^{int}, \mathbf{x}^{app}, \mathbf{x}^{loc}, \mathbf{x}^{grad}]$, then our CRF can be written as:

¹ Funding for this research was provided by Kodak research and Carestream.

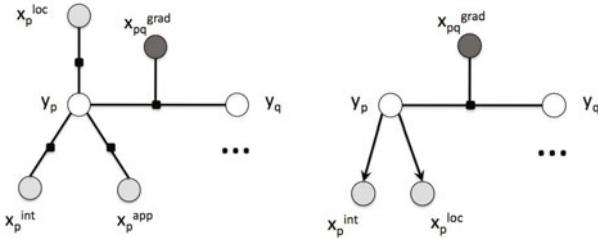


Fig. 3. The crf model and the crf-mrf graphical models explained in the text

$$\begin{aligned}
 p(\mathbf{y}|\mathbf{x}) = & \frac{1}{Z(\mathbf{x})} \prod_p \exp \left(- \sum_{c=1}^C \sum_{b=1}^B \lambda_{c,b} \rho_\lambda(x_p^{int}, y_p) - \sum_{c=1}^C \sum_{l=1}^{L_c} \alpha_{c,l} \rho_\alpha(x_p^{loc}, y_p) \right) \\
 & \prod_p \exp \left(- \sum_{c=1}^C \sum_{t=1}^T \tau_{c,t} \rho_\tau(x_p^{app}, y_p) \right) \prod_{p,q \in N} \exp \left(- \sum_{c1,c2} \sum_{g=0}^G \gamma_{g,c1,c2} \rho_\gamma(y_p, y_q, x_{pq}^{grad}) \right)
 \end{aligned} \quad (1)$$

where C indicates total number of class labels, B indicates number of bins per class (and we assume that each class has the same number of bins), T indicates number of appearance clusters per class, L_c indicates number of Gaussian distributions used to model the class c , G indicates the number of gradient bins and $Z(\mathbf{x})$ is the normalization constant. We note that we use HOG features in place of appearance features in the later experiments and the product of HOG feature potentials that is included in the equation (with $\mathbf{x} \equiv [\mathbf{x}^{int}, \mathbf{x}^{hog}, \mathbf{x}^{loc}, \mathbf{x}^{grad}]$) is as follows.

$$\prod_p \exp \left(- \sum_{c=1}^C \sum_{h=1}^H \beta_{c,h} \rho_\beta(x_p^{hog}, y_p) \right) \quad (2)$$

The parameters (in equations 1 and 2) of the data cost term are $\lambda_{c,b}$ one for each bin of each class label c , $\tau_{c,t}$ one for each appearance texton of each class label c , $\beta_{c,h}$ one for each HOG feature cluster for each class label c , $\alpha_{c,b}$ one for each location Gaussian for each class label c and the interaction parameters are $\gamma_{g,c1,c2}$, one for each gradient bin for each pair of classes. The feature functions have the following definitions.

$$\begin{aligned}
 \rho_\lambda(x_p^{int}, y_p) &= \{1 \text{ if } x_p^{int} \in \text{bin } b \text{ and } y_p = c, 0 \text{ otherwise}\} \\
 \rho_\tau(x_p^{app}, y_p) &= \{1 \text{ if } y_p = c \text{ and } x_p^{app} = t, 0 \text{ otherwise}\} \\
 \rho_\beta(x_p^{hog}, y_p) &= \{1 \text{ if } y_p = c \text{ and } x_p^{hog} = h, 0 \text{ otherwise}\} \\
 \rho_\alpha(x_p^{loc}, y_p) &= \{1 \text{ if } y_p = c \text{ and } x_p^{loc} = l, 0 \text{ otherwise}\} \\
 \rho_\gamma(y_p, y_q, x_{pq}^{grad}) &= \begin{cases} 1 & \text{when } y_p = c1 \text{ and } y_q = c2 \text{ and the gradient is in bin } g \\ 0 & \text{otherwise,} \end{cases}
 \end{aligned}$$

where $x_p^{app} = t$ indicates that the appearance feature at pixel p is closest to the texton codeword t . $x_p^{loc} = l$ indicates that the location feature at pixel p has highest probability of belonging to the Gaussian distribution $\{c, l\}$.

For our CRF-MRF model (as shown in Fig. 3) the local (datacost) features are modeled generatively while the interactive terms are modeled discriminatively. The equations of the model are given as follows:

$$\begin{aligned} p(\mathbf{x}^{\text{loc}}, \mathbf{x}^{\text{int}}, \mathbf{y} | \mathbf{x}^{\text{grad}}) &= p(\mathbf{x}^{\text{loc}} | \mathbf{y}) p(\mathbf{x}^{\text{int}} | \mathbf{y}) p(\mathbf{y} | \mathbf{x}^{\text{grad}}) \\ &= \prod_{p \in V} \mathcal{N}(x_p^{\text{loc}} | \mu_{y_p}, \Sigma_{y_p}) \prod_{p \in V} p(x_p^{\text{int}} | y_p) \frac{1}{Z(\mathbf{x}^{\text{grad}})} \prod_{p, q \in N} \exp \left(- \sum_{c1, c2} \sum_{g=0}^G \gamma_g \rho_g(y_p, y_q, x_{pq}^{\text{grad}}) \right), \end{aligned} \quad (3)$$

where $p(\mathbf{x}^{\text{loc}} | \mathbf{y})$ for location is modeled as a Gaussian mixture model and $p(\mathbf{x}^{\text{int}} | \mathbf{y})$ is modeled as a discrete distribution. Learning here is easier as the learning of the interaction parameters is not affected by the intensity, location or appearance features.

We perform maximum likelihood (ML) or conditional ML learning using standard gradient descent. In contrast to [8] and other pseudo-likelihood or autoregression approaches such as [1] we use a fully globally defined CRF. However, for learning with gradient descent we need model expectations involving intractable marginal distributions during learning due to the cyclic nature of lattices. We therefore use a form of approximate global inference known as variational message passing (VMP) [15] to obtain approximate marginals for learning and loopy max-product for final predictions and segmentations. Interestingly, when learning pairwise potentials in both MRFs and CRFs we have often found that pairwise functions can become non-submodular and thus popular alternatives for inference such as those based on graph-cuts [2] cannot be used due to their modularity constraints.

3 Experiments and Results

We identify the background, liver, right kidney, left kidney, gall bladder and spleen as **C1**, **C2**, **C3**, **C4**, **C5** and **C6** respectively. These organs were selected because they pose a challenge to the learning algorithms. For example, the intensity profiles of the spleen and liver are very similar as well as to the stomach and intestines that are included in the background class. We perform 3 different experiments.

The first two experiments are in an interactive setup where the training slices and test slices come from the same patient. The goal is to accelerate the laborious task of labeling volumes in which a few slices are labeled with some time for an offline learning phase. The user can correct the results in a subsequent session. For both of these, we use 17 patient volumes. 8 of these volumes are used to compute the location information and 9 volumes to train and test on. We use 50 bins per class for intensity and 45 bins per pair of classes for gradient pairwise bins. Location features are modeled as mixture of Gaussians and the choice of the number of Gaussians in the mixture is determined by the organ's size and compactness. We use 11 Gaussians for the background, 6 for the liver, 3 for the right kidney, 3 for the left kidney, 3 for the gall bladder and 5 for the spleen. Initial clusters are obtained from K-means clustering on sampled points (taken from a completely different set of patient volumes and not from the training or testing slices), then running a Expectation Maximization (EM) based Gaussian mixture model with diagonal covariance matrices. We use 180 appearance code words and 120 HOG codewords. We choose a pair of 2 consecutive slices (mostly so we can include all the

classes) for training and so that we can extract gradients in the 3rd (z-axis) dimension during training. For testing, we use 4 pairs of 4 consecutive slices. We can use larger number of consecutive slices at the cost of increased memory requirements.

In our first experiment, we compare the simple logistic regression with CRF and compare different features. We compare use of intensity, location (IL) with the addition of appearance (ILA) or the addition of HOG (ILH) or addition of both appearance and HOG (ILAH). Though the appearance patches 5x5 are not identical to the HOG dimensions (cell size of 8x8), we use a larger number of clusters for appearance. We observed that when we used appearance patches of 9x9, the segmentation results were blockier and hence worse. We initialize the datacost CRF parameters using the corresponding

Table 1. This table compares logistic regression with CRFs and also compares different features. I, L, A and H stand for intensity, location, appearance and HOG respectively. ACA and PA stands for average class accuracy and pixel accuracy respectively. All values are percentage accuracy.

(a) Experiment 1

	logreg				CRF			
	IL	ILA	ILH	ILAH	IL	ILA	ILH	ILAH
C1	96.2	95.7	97.5	95.6	98.2	97.3	98.4	96.4
C2	90.8	78.2	91.3	81.2	90.5	79.3	93.6	81.6
C3	67.4	55.6	75.0	51.4	72.9	64.0	80.5	57.7
C4	68.7	65.9	75.0	56.6	75.4	69.3	78.6	61.7
C5	19.2	7.5	40.3	8.0	19.7	12.7	42.4	21.8
C6	85.5	56.2	86.5	60.9	89.3	62.2	90.3	71.6
ACA	71.3	59.8	77.6	59.0	74.4	64.1	80.6	65.1
PA	92.9	88.9	94.5	89.1	94.9	90.9	95.9	90.7

(b) Experiment 3

	logreg			CRF		
	IL	IH	ILH	IL	IH	ILH
82.5	83.8	83.1	83.2	83.8	85.7	
67.8	78.0	75.7	68.6	78.0	74.9	
63.3	39.1	66.9	63.3	39.1	64.9	
82.8	42.5	85.5	82.8	42.5	83.8	
97.9	50.2	89.7	97.7	50.2	91.9	
84.6	67.3	83.8	84.6	67.3	78.2	
79.8	60.1	80.7	80.0	60.1	80.0	
79.9	81.0	81.6	80.6	81.0	83.4	

logistic regression parameters. The results are shown in Table. 1(a). In the second experiment, we compare the CRF to the CRF-MRF. We use only intensity and location to compare the performance of these two approaches. Results are shown in Table. 2.

In the third experiment, we use all 22 patient volumes in a typical training/testing framework. We leave out 3 volumes for testing. Of the remaining 19 volumes, we use 15 volumes for training and 4 for cross-validation to select different modeling and parameter settings. We use the full 3D volume of patient for training and inference. The logistic regression is modified to handle imbalance of classes (Appendix A)². The datacost parameters learnt here are used in the CRF and kept fixed. The smoothness term parameters are learnt in the CRF. We use 100 bins per class for intensity and 45 bins per pair of classes for gradient pairwise bins. Location features are modeled as mixture of Gaussians and the choice of the number of Gaussians in the mixture is determined by validation set likelihood and we note that the results tend to depend on the organ's size and compactness. The best set of Gaussians were 30 for the background, 20 for the liver, 5 for the right kidney, 3 for the left kidney, 2 for the gall bladder and 7 for the spleen. A weighted form of Expectation Maximization (EM) (Appendix B) based Gaussian mixture model with full covariance matrices was used to allow using more data points. 800

² <http://www.cs.rochester.edu/~bhole/medicalseg>

Table 2. The tables above are the confusion matrices obtained using CRF (left) and CRF-MRF (right). The zeros are left out of the confusion matrices for clarity.

		Predicted (CRF)								Predicted (CRF-MRF)					
		C1	C2	C3	C4	C5	C6			C1	C2	C3	C4	C5	C6
Actual	C1	98.2	0.9	0.2	0.2	0.1	0.2			98.1	1.1		0.2	0.6	
	C2	9.0	90.5	0	0.2	0.2	0.1			44.2	52.4		2.3	1.1	
	C3	22.0	0.0	72.9			5.1			100.0					
	C4	15.9	8.5	0.2	75.4		0			93.7	0.4		6.0	0.0	
	C5	69.9	10.4			19.7	0			54.8	6.2		0.2	38.8	
	C6	8.1	1.1	1.4	0.0	0	89.3			99.9	0.1				
	ACA = 74.36 , PA = 94.91								ACA = 32.54 , PA = 81.8						

HOG codewords were generated using K-means. More number of codewords belonged to larger classes. The results are shown in Table 1(b).

4 Discussion and Conclusions

The experiments show an increase in pixel and class accuracy when HOG is used in most of the cases. The interactive setup experiment shows the use of a discriminative structured model like the CRF doing better than the simpler logistic regression. The appearance features do not tend to do well and it is possible that more appearance clusters or code words are required. In the 3rd experiment, we note that use of the modified logistic regression improved average class accuracy without which classes like gall bladder and kidney have very large error rates. We also note that in our setting, CRFs improve the pixel accuracy to some extent while the class accuracy remains almost the same. Using intensity and HOG (IH) which is less reliant on the coarse registration step shows promise. The class accuracy increases from 30% (only intensity) to 60% (IH) (table not shown). The location feature however is more dependent on the manual registration step. The second experiment shows improved performance when using the discriminative CRF model compared to the MRF-CRF even with simpler features. CRF-MRF model completely misses some organs like the spleen (C6). In general, the goal is to reduce the labeling time of the user and so training can be done offline in this setup so the user can make corrections in subsequent interactions. Future work will involve use of shape models along with discriminative models to improve segmentations further.

References

1. Blake, A., Rother, C., Brown, M., Perez, P., Torr, P.: Interactive image segmentation using an adaptive GMMRF model. In: Pajdla, T., Matas, J(G.) (eds.) ECCV 2004. LNCS, vol. 3021, pp. 428–441. Springer, Heidelberg (2004)
2. Boykov, Y., Veksler, O., Zabih, R.: Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 23 (2001)
3. Criminisi, A., Shotton, J., Robertson, D.P., Konukoglu, E.: Regression forests for efficient anatomy detection and localization in CT studies. In: MCV, pp. 106–117 (2010)

4. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE CVPR, pp. 886–893. IEEE Computer Society, Washington, DC, USA (2005)
5. Graf, F., Kriegel, H.P., Schubert, M., Strukelj, M., Cavallaro, A.: Fully automatic detection of the vertebrae in 2d ct images. In: SPIE Medical Imaging, vol. 7962 (2011)
6. Kumar, S., Hebert, M.: Discriminative random fields: A discriminative framework for contextual interaction in classification. In: ICCV, pp. 1150–1157 (2003)
7. Lafferty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data (2001)
8. Lee, C.H., Wang, S., Murtha, A., Brown, M.R.G., Greiner, R.: Segmenting brain tumors using pseudo-conditional random fields. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part I. LNCS, vol. 5241, pp. 359–366. Springer, Heidelberg (2008)
9. Ling, H., et al.: Hierarchical, learning-based automatic liver segmentation. In: IEEE CVPR (2008)
10. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. IJCV 60(2) (2004)
11. Motwani, K., Adluru, N., Hinrichs, C., Alexander, A.L., Singh, V.: Epitome driven 3-d diffusion tensor image segmentation: on extracting specific structures. In: NIPS (2010)
12. Seifert, S., et al.: Hier. parsing and semantic nav. of full body CT data. In: Proc. SPIE (2009)
13. Tsechpenakis, G., Wang, J., Mayer, B., Metaxas, D.: Coupling CRFs and deformable models for 3D medical image segmentation, pp. 1–8 (2007)
14. Varshney, L.: Abdominal organ segmentation in ct scan images: A survey (2002)
15. Winn, J., Bishop, C.M.: Variational message passing. J. Mach. Learn. Res. 6, 661–694 (2005)
16. Zhang, Y., Brady, M., Smith, S.: Segmentation of brain MR images through a hidden markov random field model and the EM algorithm. IEEE Trans. Med. Imaging 20(1), 45–57 (2001)

Automatic Human Knee Cartilage Segmentation from Multi-contrast MR Images Using Extreme Learning Machines and Discriminative Random Fields

Kunlei Zhang and Wenmiao Lu

School of Electrical & Electronic Engineering, Nanyang Technological University

Abstract. Accurate and automatic segmentation of knee cartilage is required for the quantitative cartilage measures and is crucial for the assessment of acute injury or osteoarthritis. Unfortunately, the current works are still unsatisfactory. In this paper, we present a novel solution toward the automatic cartilage segmentation from multi-contrast magnetic resonance (MR) images using a pixel classification approach. Most of the previous classification based works for cartilage segmentation only rely on the labeling by a trained classifier, such as support vector machines (SVM) or k-nearest neighbor, but they do not consider the spatial interaction. Extreme learning machines (ELM) have been proposed as the training algorithm for the generalized single-hidden layer feedforward networks, which can be used in various regression and classification applications. Works on ELM have shown that ELM for classification not only tends to achieve good generalization performance, but also is easy to be implemented since ELM requires less human intervention (only one user-specified parameter needs to be chosen) and can get direct least-square solution. To incorporate spatial dependency in classification, we propose a new segmentation method based on the convex optimization of an ELM-based association potential and a discriminative random fields (DRF) based interaction potential for segmenting cartilage automatically with multi-contrast MR images. Our method not only benefits from the good generalization classification performance of ELM but also incorporates the spatial dependencies in classification. We test the proposed method on multi-contrast MR datasets acquired from 11 subjects. Experimental results show that our method outperforms the classifiers based solely on DRF, SVM or ELM in segmentation accuracy.

1 Introduction

Magnetic resonance (MR) imaging has emerged as the most promising imaging modality to detect structural changes in cartilages, as it can provide direct and noninvasive images of the whole knee joint. Since the assessment of acute injury or osteoarthritis (OA) requires exact quantitative cartilage measures such as the cartilage's thickness and volume, accurate cartilage segmentation as the key to these measures has gained considerable attention in recent years. In general, Manual and semi-automatic segmentation is laborious and time consuming. Moreover, they are prone to inter/intra-observer variability rendering the analysis of the results very complicated. Therefore, it is

desirable to automate the segmentation. Unfortunately, automatic cartilage segmentation from MR images is a challenging task due to the thin variable morphology of cartilages, intensity inhomogeneity, the low contrast between cartilages and other soft tissues, and MR artifacts. In this work, we address the problem of automatic human knee cartilage segmentation from multiple sets of MR images taken with different sequences (referred to as *multi-contrast MR images*). To the best of our knowledge, the research on cartilage segmentation from multi-contrast MR data is sparse.

1.1 Related Works

In recent years, several automatic approaches to cartilage segmentation have been proposed. Glocker et al. [1] utilized a nonrigid registration scheme to segment the patellar cartilage by fitting a statistical atlas to MR data. Fripp et al. [2] used a hierarchical cartilage segmentation scheme based on a hybrid deformable model. Dodin et al. [3] developed a hierarchical automatic segmentation algorithm for knee cartilage volume quantification. At the 2010 MICCAI conference, a competition for knee segmentation was held at a workshop “Medical Image Analysis for the Clinic” [4]. Among these automatic segmentation methods, model-based works ranked in the highest quantile, such as the approach proposed in [5]. Besides, image segmentation can also be considered as a statistical classification problem in which each pixel belongs to a class. Folkesson et al. [6] employed a two step k-nearest neighbor classifier to automatically separate cartilages from non-cartilages. While all the aforementioned works are based on a single MR sequence, multi-contrast MR images provide different contrast mechanisms between tissues and help separate different tissues. Koo et al. [7] proposed to segment cartilage automatically with multi-contrast MR images using support vector machines (SVM). However, these classification based works [6, 7] assumed that data instances were independent, which may not be appropriate for the cartilage segmentation task. As pointed out by [8], class labels are not independent in most real-world spatial classification problems, where correlations in labels of neighboring instances exist in data with multi-dimensional structure, such as images and volumes. This motivates one to incorporate contextual information in the form of spatial dependencies in the classification for cartilage segmentation, which generally yields smoother and more reliable results.

1.2 Overview of the Work Presented

Extreme learning machines (ELM) [9, 10] were proposed as the training algorithm for the generalized single-hidden layer feedforward networks (SLFN), which can be used in various regression and classification applications. Works on ELM have shown that ELM for classification not only tends to achieve good generalization performance, but also is easy to be implemented since ELM requires less human intervention (only one user-specified parameter needs to be chosen) and can get direct least-square solution. Unfortunately, ELM does not consider spatial information which is necessary and beneficial to the cartilage segmentation. Probabilistic graphical models such as discriminative random fields (DRF) have been used to incorporate spatial contextual constraints in many applications [8].

In this paper, we present a novel solution to the problem of automatic cartilage segmentation with multi-contrast MR images based on ELM and DRF. The proposed method not only benefits from the good generalization performance of the classification based on ELM but also incorporates the spatial dependencies in classification. Specifically, we adopt the classification framework based on the convex optimization of an ELM-based association potential and a DRF-based interaction potential for cartilage segmentation. We also employ a feature set encoding diverse forms of image and anatomical structure information, including intensity values and geometrical information which are crucial to distinguish cartilages from other tissues. We finally perform the loopy belief propagation (LBP) inference algorithm to find the optimal label configuration. The experimental results of applying our method on multi-contrast knee MR data show the improvements on segmenting cartilage over other classification approaches including DRF, SVM and ELM.

2 Methodology

Throughout this paper, let $\mathbf{x} = \{\mathbf{x}_i\}_{i \in S}$ denote the observed data, where the observation \mathbf{x}_i of the i th instance is represented by d -dimensional feature vector, and S is the set of instances. The corresponding labels are given by $\mathbf{y} = \{y_i\}_{i \in S}$, where y_i is the class label of the i th instance. The aim of the classification based segmentation is to infer the most likely joint class labels $\mathbf{y}^* = \{y_i^*\}_{i \in S}$ based on \mathbf{x} by a classification model.

2.1 Conditional and Discriminative Random Fields

Conditional random fields (CRF) [11] are a discriminative approach which directly models the posterior distribution $p(\mathbf{y} | \mathbf{x})$ without building the joint distribution. CRF relaxes the conditional independence assumption of the observations and allows the modeling of complex dependencies (1) between the label of an instance and its features, (2) between the labels of adjacent instances, and (3) between the labels of adjacent instances and their features.

DRF [8] was proposed as a multi-dimensional extension of CRF for lattice-structured data. The DRF model can be formulated as

$$p(\mathbf{y} | \mathbf{x}) = \frac{1}{Z} \prod_{i \in S} A(y_i, \mathbf{x}) \prod_{i \in S} \prod_{j \in N_i} I(y_i, y_j, \mathbf{x}) \quad (1)$$

where Z is a normalization constant called the partition function, $A(y_i, \mathbf{x})$ is the association potential modeling dependencies between y_i and \mathbf{x} , and $I(y_i, y_j, \mathbf{x})$ is the interaction potential modeling dependencies between y_i, y_j and \mathbf{x} . DRF is a powerful method for modeling dependencies in spatial data, but the logistic regression for the association potential [8] often cannot estimate appropriate parameters for the problems with unbalanced class labels, high-dimensional feature spaces, or highly correlated features. Due to this, in some tasks DRF is not able to produce results as accurate as powerful classification models such as ELM and SVM.

2.2 Extreme Learning Machines

The main feature of ELM [9, 10] is that the hidden layer parameters are randomly generated without iterative tuning. After randomly generating L hidden nodes, the outputs of the L hidden nodes with respect to the input \mathbf{x} are presented as $\mathbf{h}(\mathbf{x}) = [h_1(\mathbf{x}), \dots, h_L(\mathbf{x})]$. $\mathbf{h}(\mathbf{x})$ actually maps the data from the d -dimensional input space to the L -dimensional hidden layer feature space. Let $\boldsymbol{\beta} = [\beta_1, \dots, \beta_L]^T$ be the vector of the output weights between the L hidden nodes and the output nodes. The output function of ELM for generalized SLFN is

$$f(\mathbf{x}) = \sum_{i=1}^L \beta_i h_i(\mathbf{x}) = \mathbf{h}(\mathbf{x}) \boldsymbol{\beta} \quad (2)$$

The minimal norm least square solution can be obtained

$$\boldsymbol{\beta} = \mathbf{H}^T \mathbf{T} \quad (3)$$

where \mathbf{T} is the label matrix, and \mathbf{H}^T is the Moore-Penrose generalized inverse of the hidden layer output matrix $\mathbf{H} = [\mathbf{h}^T(\mathbf{x}_1), \dots, \mathbf{h}^T(\mathbf{x}_N)]^T$. One of the methods to calculate Moore-Penrose generalized inverse of a matrix is the orthogonal projection method: $\mathbf{H}^T = (\mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T$. According to the ridge regression theory, one can add a user-specified positive value k to the diagonal of $\mathbf{H}^T \mathbf{H}$ and the resultant solution in the following is more stable and tends to have better generalization performance.

$$f(\mathbf{x}) = \mathbf{h}(\mathbf{x}) \boldsymbol{\beta} = \mathbf{h}(\mathbf{x}) (k \mathbf{I} + \mathbf{H}^T \mathbf{H})^{-1} \mathbf{H}^T \mathbf{T} \quad (4)$$

The works of [9, 10] indicate that ELM can achieve good generalization performance as long as the number of hidden nodes L is large enough, that is, the performance of ELM is not sensitive to L and L thus needs not to be tuned. Therefore, only one parameter k in (4) needs to be chosen in ELM instead two or more user-specified parameters in SVM and less human intervention is required. Furthermore, seen from (4), ELM can get direct least-square solution and avoids quadratic programming in SVM. Hence, the implementation of ELM can be easier than the traditional SVM.

2.3 The Proposed Method

A. Classification Model

Although ELM tends to achieve good generalization performance and can be easier to implement than SVM in many classification applications, ELM assumes that data instances are independent and identically distributed (i.i.d), which does not consider interactions on the labels of adjacent data instances and may produce undesirable results for the cartilage segmentation task. Conversely, DRF considers these interactions but does not have the appealing generalization ability as ELM and SVM. To incorporate spatial dependencies and model contextual interactions in the classification process of ELM, we present a unified formulation of ELM and DRF as follows

$$p(\mathbf{y} | \mathbf{x}) = \frac{1}{Z} \prod_{i \in S} \frac{1}{1 + e^{-y_i (w_0 + w_i f(\gamma_i(\mathbf{x}))}}} \prod_{i \in S} \prod_{j \in N_i} e^{y_j y_i v^T \gamma_{ij}(\mathbf{x})} \quad (5)$$

where we get the association potential by converting the output of the ELM decision function $f(\cdot)$ to a posterior probability using the sigmoid function and in the inter-action potential we use a DRF model to incorporate the interactions in labels and make it data-adaptive. $\mathbf{w} = [\mathbf{w}_0, \mathbf{w}_1]^T$ and \mathbf{v} are the association potential and interaction potential parameters, respectively, to be estimated in the training stage. $\gamma_i(\mathbf{x})$ is the feature vector for node i , while $\gamma_{ij}(\mathbf{x})$ is the feature vector for a pair of neighboring nodes i and j computed from the observations \mathbf{x} . $\gamma_{ij}(\mathbf{x})$ can be set by taking the absolute difference of $\gamma_i(\mathbf{x})$ and $\gamma_j(\mathbf{x})$, that is, $\gamma_{ij}(\mathbf{x}) = |\gamma_i(\mathbf{x}) - \gamma_j(\mathbf{x})|$ which penalizes for high absolute differences between the features of neighboring nodes.

B. Features

Features based solely on intensity values are insufficient to extract cartilage due to overlapping intensity distributions and ambiguous boundaries between the cartilage and non-cartilages. Fortunately, the geometrical information of the anatomical structure in the multi-contrast MR images has proven to be crucial to distinguish cartilage from surrounding tissues [7]. The geometrical information is obtained from automatically segmenting the bones, which can be robustly done by thresholding the multi-contrast images. Subsequently, the centers in both of the distal femoral condyles and the femoral centerlines are computed. Finally, Euclidean distance from the closest bone and gradient of the distance, angle between the main magnetic direction of the MRI and the line to the femoral center line, and relative location along the medial and lateral center of the distal femur are calculated as the geometrical information based features for each image pixel. Thus, the feature set adopted in this paper consists of normalized intensity values of multi-contrast MR images as well as geometrical information based features.

C. Parameter Learning and Inference

In the classification model (5), we first compute the ELM decision function $f(\cdot)$ by solving a least-square problem as in [9]. Then the parameters $\Theta = \{\mathbf{w}, \mathbf{v}\}$ are simultaneously estimated with M training images using pseudolikelihood,

$$\Theta^* = \arg \max_{\Theta} \prod_{m=1}^M \prod_{i \in S} p(y_i^m | y_{N_i}^m, \mathbf{x}^m, \mathbf{w}, \mathbf{v}) \quad (6)$$

where

$$p(y_i^m | y_{N_i}^m, \mathbf{x}^m, \Theta) = \frac{1}{z_i^m} \frac{1}{1 + e^{-y_i^m \mathbf{w}^T \mathbf{f}(\gamma_i(\mathbf{x}^m))}} \prod_{j \in N_i} e^{y_i^m y_j^m \mathbf{v}^T \gamma_{ij}(\mathbf{x}^m)}$$

with

$$\mathbf{f}(\gamma_i(\mathbf{x}^m)) = [1 \ f(\gamma_i(\mathbf{x}^m))]^T$$

and

$$z_i^m = \sum_{y_i^m \in \{-1, 1\}} \left(\frac{1}{1 + e^{-y_i^m \mathbf{w}^T \mathbf{f}(\gamma_i(\mathbf{x}^m))}} \prod_{j \in N_i} e^{y_i^m y_j^m \mathbf{v}^T \gamma_{ij}(\mathbf{x}^m)} \right)$$

To prevent over-fitting, we utilize the L2-regularization, which involves adding a penalty term in the form of a sum of squares of all the parameters in order to

discourage the parameters from reaching large values. We thus have the following penalized negative logarithm pseudolikelihood

$$\Theta^* = \arg \min_{\Theta} \left[\sum_{m=1}^M \sum_{i \in S} (\log(z_i^m) - y_i^m \mathbf{w}^T \mathbf{f}(\gamma_i(\mathbf{x}^m)) - \sum_{j \in N_i} y_i^m y_j^m \mathbf{v}^T \gamma_j(\mathbf{x}^m)) + \lambda_1 \mathbf{w}^T \mathbf{w} + \lambda_2 \mathbf{v}^T \mathbf{v} \right] \quad (7)$$

where λ_1 and λ_2 are nonnegative regularizing constants determined by cross-validation. Since the penalized logarithm pseudo-likelihood in (7) is jointly convex with respect to the parameters Θ , it can be easily minimized using gradient descent.

For a given new test knee MR image \mathbf{x}' , the inference problem is to find an optimal label configuration \mathbf{y}' based on the learned model parameters Θ^* . We perform the LBP inference algorithm [12] to solve this problem. In the implementation, we utilize the UGM toolbox provided by Dr. Mark Schmidt [13].

3 Experiments and Results

3.1 Description of Multi-contrast Human Knee MR Images and Preprocessing

We validate our cartilage segmentation method on multi-contrast MR datasets acquired from 11 volunteers who were not known to have OA or knee pain. Multiple sets of MR images were taken for each knee joint using a 3.0T magnet scanner (GE Healthcare, Waukesha, WI) with multiple MR sequences including FS SPGR, FIESTA, and IDEAL GRE. Different MR sequences were acquired using different sets of parameters. FS SPGR sequences were obtained with echo time (TE) 4ms, repetition time (TR) 28ms, and flip angle (FA) 25°; FIESTA with TE 3.5ms, TR 7.2ms and FA 25°; IDEAL GRE (water & fat) with TE 3.4ms, TR 9.4ms and FA 30°. In addition, several parameters were common: all images acquired in the sagittal plane with slice thickness 1.5mm, in-plane spacing 0.625mm×0.625mm and matrix size 256×256. The golden standard cartilage segmentation for each knee was obtained by an expert's manual segmentation in the form of binary mask images of cartilage.

Preprocessing is required before segmenting cartilage from these knee MR data. The multiple sets of MR images are aligned. We removed regions that contain no cartilage information and the volume size is reduced to 150×188×55. Intensities of all sets of MR images were normalized to [0, 1], which brings them within the same dynamic range, improving the stability of the classifier.

3.2 Evaluation Metrics

Besides visual evaluation, the cartilage segmentations automatically obtained are quantitatively compared to the golden standard using *sensitivity*, *specificity* and Dice similarity coefficient (*DSC*) as used in existing cartilage segmentation works [1, 2, 3, 6, 7]. $Sensitivity = TP / (TP + FN)$ is the true positive fraction, $specificity = TN / (FP + TN)$ is the true negative fraction and $DSC = 2TP / (2TP + FN + FP)$ is a spatial overlap index, where *TP* is true positive, *TN* is true negative, *FP* is false positive, and *FN* is false negative counts for the pixels.

3.3 Results

As in *many works* on the problem of segmenting medical images [14, 15], we employ a subject-specific manner: for each subject (knee), the classifiers are trained on a subset of the subject's data and then tested on another (disjoint) subset. We use every fifth slice as the training data and other slices as the testing data. In the experiments, we utilize the sigmoid additive hidden nodes $\mathbf{h}(\mathbf{x}) = [G(\mathbf{a}_1, \mathbf{b}_1, \mathbf{x}), \dots, G(\mathbf{a}_L, \mathbf{b}_L, \mathbf{x})]$ with $G(\mathbf{a}, \mathbf{b}, \mathbf{x}) = 1/(1 + \exp(-(\mathbf{a}^T \mathbf{x} + b)))$ in ELM. All the hidden node parameters $\{\mathbf{a}_i, \mathbf{b}_i\}_{i=1}^L$ are randomly generated. The number of hidden node is set to be $L = 5000$ and the parameter k in (4) is chosen from a wide range $\{2^{-20}, 2^{-18}, \dots, 2^{18}, 2^{20}\}$. λ_1 and λ_2 in (7) are chosen from the range $\{5 \times 10^{-9}, 5 \times 10^{-7}, \dots, 5 \times 10^7, 5 \times 10^9\}$ by cross-validation.

A. Qualitative Results

We compare our method against DRF, SVM, and ELM. In Figure 1, we illustrate the qualitative results for *one slice of a knee joint* using the four classifiers as an example. The segmentations of SVM and ELM (Figure 1 (d), (e)) show high and scattered FN,

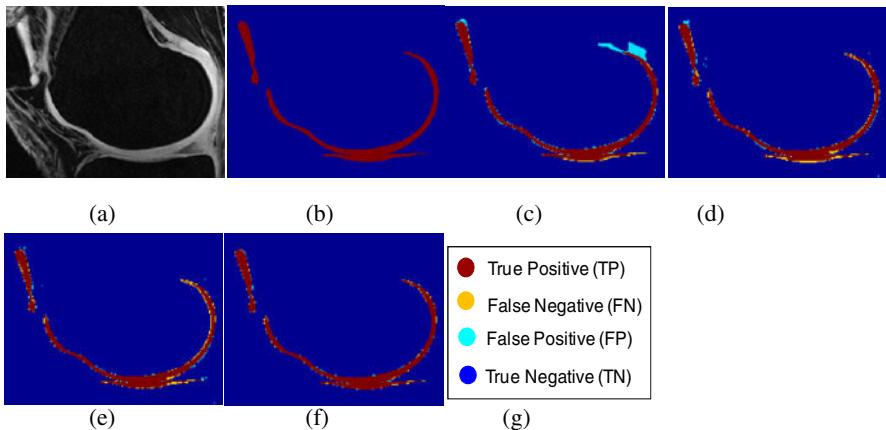


Fig. 1. One example of the segmentation results. (a) the cropped MR image for one slice of one knee, (b) the gold standard, (c), (d), (e), and (f) segmentation results using DRF, SVM, ELM and our method respectively, (g) the legend.

Table 1. Mean \pm std of sensitivity, specificity, and DSC for all slices of all knee joints using DRF, SVM, ELM and our method, respectively

Method	Quantitative Measures		
	Sensitivity	Specificity	DSC
DRF	0.836 \pm 0.098	0.972 \pm 0.024	0.764 \pm 0.173
SVM	0.871 \pm 0.116	0.996 \pm 0.003	0.863 \pm 0.114
ELM	0.875 \pm 0.117	0.997 \pm 0.003	0.870 \pm 0.120
Our method	0.913\pm0.092	0.997\pm0.002	0.909\pm0.100

which is because both SVM and ELM assume image pixels are i.i.d and ignore the contextual interactions. The segmentation of DRF in Figure 1(c) gives high as well as dense *FP* and *FN*, which is because DRF incorporates spatial dependencies but lacks generalization ability. Our method provides relatively low *FP* and *FN* (Figure 1(f)), because the unified model of ELM and DRF not only incorporates the spatial correlations among neighboring image pixels but also benefits from the generalization classification ability obtained from ELM.

B. Quantitative Results

To compare segmentation performances of all the four methods quantitatively, Table 1 provides average sensitivity, specificity and DSC values for *all slices of all knee joints*. This table shows the superiority of combining DRF with ELM to using merely DRF or ELM as the classifier. Based on these results, we conclude that the proposed method provides improved cartilage segmentation results compared to other classification approaches.

4 Conclusions

In this paper, we presented a new automatic cartilage segmentation technique working on multi-contrast MR images using pixel classification. Compared with commonly used classification approaches based solely on DRF, SVM or ELM, the advantages of the proposed technique are achieved via effective incorporation of high-dimensional feature set with the ELM-based classification and the DRF model. Unlike atlas or deformable-model based segmentation methods which require training datasets from both healthy subjects and patients with OA, the proposed technique does not rely on prior information related to the pathological state of the knee. Instead, the multi-contrast MR images and the spatial dependences between neighboring soft tissues are exploited to achieve much robust cartilage segmentation from both healthy and pathological knees. We demonstrate with a comprehensive multi-contrast knee dataset that the proposed method outperforms the segmentation techniques based solely on DRF, SVM or ELM in accuracy.

References

1. Glocker, B., Komodakis, N., Paragios, N., Glaser, C., Tziritas, G., Navab, N.: Primal/Dual Linear Programming and Statistical Atlases for Cartilage Segmentation. In: Ayache, N., Ourselin, S., Maeder, A. (eds.) MICCAI 2007, Part II. LNCS, vol. 4792, pp. 536–543. Springer, Heidelberg (2007)
2. Fripp, J., Crozier, S., Warfield, S.K., Ourselin, S.: Automatic Segmentation and Quantitative Analysis of the Articular Cartilages from Magnetic Resonance Images of the Knee. IEEE Trans. Med. Imag. 29(1), 55–64 (2010)
3. Dodin, P., Pelletier, J., Martel-Pelletier, J., Abram, F.: Automatic Human Knee Cartilage Segmentation from 3-D Magnetic Resonance Images. IEEE Trans. Biomed. Eng. 57(11), 2699–2711 (2010)
4. Heimann, T., Morrison, B.J., Styner, M.A., Niethammer, M., Warfield, S.K.: Segmentation of Knee Images: A Grand Challenge. In: Proc. Medical Image Analysis for the Clinic: A Grand Challenge, pp. 207–214 (2010)

5. Vincent, G., Wolstenholme, C., Scott, I., Bowes, M.: Fully Automatic Segmentation of the Knee Joint using Active Appearance Models. In: Proc. Medical Image Analysis for the Clinic: A Grand Challenge, pp. 224–230 (2010)
6. Folkesson, J., Dam, E.B., Olsen, O.F., Pettersen, P.C., Christiansen, C.: Segmenting Articular Cartilage Automatically Using a Voxel Classification Approach. IEEE Trans. Med. Imag. 26(1), 106–115 (2007)
7. Koo, S., Alto, P., Hargreaves, B.A., Gold, G.E.: Automatic Segmentation of Articular Cartilage from MRI. Patent, US 2009/0306496 (2009)
8. Kumar, S., Hebert, M.: Discriminative Random Fields. Int. J. of Comp. Vision 68(2), 179–201 (2006)
9. Huang, G.B., Zhu, Q.Y., Siew, C.K.: Extreme Learning Machine: Theory and Applications. Neurocomputing 70, 489–501 (2006)
10. Huang, G.B., Wang, D., Lan, Y.: Extreme Learning Machines: A Survey. Int. J. of Machine Learning and Cybernetics 2(2), 107–122 (2011), ELM website, <http://www.extreme-learning-machines.org/>
11. Lafferty, J., Pereira, F., McCallum, A.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proc. Int. Conf. on Machine Learning, pp. 282–289 (2001)
12. Murphy, K.P., Weiss, Y., Jordan, M.I.: Loopy Belief Propagation for Approximate Inference: an Empirical Study. In: Foo, N.Y. (ed.) AI 1999. LNCS, vol. 1747, pp. 467–475. Springer, Heidelberg (1999)
13. UGM Toolbox, <http://people.cs.ubc.ca/~schmidtm/Software/UGM.html>
14. García, C., Moreno, J.A.: Kernel Based Method for Segmentation and Modeling of Magnetic Resonance Images. In: Lemaître, C., Reyes, C.A., González, J.A. (eds.) IBERAMIA 2004. LNCS (LNAI), vol. 3315, pp. 636–645. Springer, Heidelberg (2004)
15. Lee, C., Wang, S., Brown, M., Murtha, A., Greiner, R.: Segmenting Brain Tumors Using Pseudo-Conditional Random Fields. In: Metaxas, D., Axel, L., Fichtinger, G., Székely, G. (eds.) MICCAI 2008, Part I. LNCS, vol. 5241, pp. 359–366. Springer, Heidelberg (2008)

MultiCost: Multi-stage Cost-sensitive Classification of Alzheimer's Disease

Daoqiang Zhang^{1,2} and Dinggang Shen¹

¹ Dept. of Radiology and BRIC, University of North Carolina at Chapel Hill, NC 27599

² Dept. of Computer Science and Engineering, Nanjing University of Aeronautics and Astronautics, Nanjing 210016, China

{zhangd, dgshen}@med.unc.edu

Abstract. Most traditional classification methods for Alzheimer's disease (AD) aim to obtain a high accuracy, or equivalently a low classification error rate, which implicitly assumes that the losses of all misclassifications are the same. However, in practical AD diagnosis, the losses of misclassifying healthy subjects and AD patients are usually very different. For example, it may be troublesome if a healthy subject is misclassified as AD, but it could result in a more serious consequence if an AD patient is misclassified as healthy subject. In this paper, we propose a multi-stage cost-sensitive approach for AD classification via multimodal imaging data and CSF biomarkers. Our approach contains three key components: (1) a cost-sensitive feature selection which can select more AD-related brain regions by using different costs for different misclassifications in the feature selection stage, (2) a multimodal data fusion which effectively fuses data from MRI, PET and CSF biomarkers based on multiple kernels combination, and (3) a cost-sensitive classifier construction which further reduces the overall misclassification loss through a threshold-moving strategy. Experimental results on ADNI dataset show that the proposed approach can significantly reduce the cost of misclassification and simultaneously improve the sensitivity, under the same or even higher classification accuracy compared with conventional methods.

Keywords: Cost-sensitive classification, cost-sensitive feature selection, MultiCost, multi-modality, Alzheimer's disease (AD).

1 Introduction

Alzheimer's disease (AD) is the most common form of dementia in elderly people worldwide. It is reported that the number of affected patients is expected to be doubled in the next 20 years [1]. Early diagnosis of AD and its prodromal stage, i.e., mild cognitive impairment (MCI), is very important for possible delay of the disease. At present, several biomarkers have been proved to be sensitive to AD, including brain atrophy measured by magnetic resonance imaging (MRI), hypometabolism measured by functional imaging (e.g., positron emission tomography (PET)), and quantification of specific proteins measured through cerebrospinal fluid (CSF) [2-3]. Over the past decades, a lot of AD classification methods have been developed based

on one or two imaging modalities [4-6]. Recently, several studies have indicated that different biomarkers provide complementary information, which may be useful for diagnosis of AD when used together [2-3]. Motivated by this finding, some recent methods have been proposed to combine multimodal biomarkers to improve classification accuracy, which can generally do better than using only a single type of biomarkers [7-8].

Existing methods for AD classification aim to correctly assign subjects to one of several classes, typically two classes, i.e., AD patients and healthy subjects. Accordingly, most of the currently available algorithms for AD classification are designed to minimize the *zero-one loss* or *error rate*, i.e., the number of misclassified subjects. This implicitly assumes that all errors are equally costly, i.e., the loss of misclassifying an AD patient as healthy is the same as that of misclassifying a healthy subject as AD. However, this assumption rarely holds in real AD diagnosis. For example, it may be troublesome if a healthy subject is misclassified as AD, but it could result in a more serious consequence or even loss of life if an AD patient is misclassified as healthy. Apparently, the misclassification cost of AD is much higher than that of a healthy subject, and this important *a prior* knowledge should be used to guide AD classification. However, to the best of our knowledge, no previous studies explicitly consider different losses of misclassifying AD patients and healthy subjects in AD diagnosis.

On the other hand, cost-sensitive learning which can deal with the classification problem with unequal costs has been studied for years in the machine learning and data mining communities, and a lot of cost-sensitive learning algorithms have been proposed [9-10]. However, directly applying the existing cost-sensitive learning algorithms for AD classification may encounter several challenges. First, data in AD classification are usually available in multiple modalities, e.g., MRI, PET and CSF, etc., while most conventional cost-sensitive learning algorithms cannot be directly applied to the multimodal data. Second, because of the high dimensionality of neuroimaging data, the AD diagnosis method generally employs a feature selection stage before the classification stage; thus, it would be more helpful to use the cost information in both stages. However, the existing cost-sensitive learning algorithms are mainly used for classification rather than feature selection. To our best knowledge, this type of study on cost-sensitive feature selection was not done previously.

In this paper, we propose a **Multi-stage Cost**-sensitive approach (**MultiCost**) for AD (or MCI) classification. The MultiCost method contains three main components: (1) a cost-sensitive feature selection which can select more AD-related brain regions by using different costs of misclassification in the feature selection stage, (2) a multimodal data fusion which effectively fuses data from MRI, PET and CSF biomarkers based on multiple kernels combination, and (3) a cost-sensitive classifier construction which further reduces the overall misclassification loss through a threshold-moving strategy. Experimental results on ADNI dataset are presented to show the efficacy of the proposed approach.

The rest of this paper is organized as follows. In Section 2, we present the proposed MultiCost method for multimodal AD/MCI classification. Experimental results are given in Section 3. Finally, Section 4 concludes this paper and indicates points for future work.

2 MultiCost

In this section, we present a Multi-stage Cost-sensitive classification approach (MultiCost) for multimodal AD or MCI classification. We call it as MultiCost because the cost information is used in multiple stages, i.e., both feature selection and classification stages. There are three main steps in MultiCost, i.e., cost-sensitive feature selection on high-dimensional brain imaging data, multimodal data fusion from MRI, PET, and CSF biomarkers, and cost-sensitive classification.

2.1 Cost-sensitive Feature Selection

Because of the high dimensionality of brain imaging data, feature selection is usually required before classification. Like cost-sensitive classification, we can also explicitly exploit different costs of misclassifying diseased patients and healthy subjects for better feature selection, which is expected to select more ‘good’ features for predicting AD patients rather than healthy subjects. Here, we focus on binary classification, i.e., diseased (‘D’) or healthy (‘H’) class. Let C_{DH} and C_{HD} denote the cost of misclassifying a diseased patient (AD or MCI) as healthy and the cost of misclassifying a healthy subject as diseased, respectively. Without loss of generality, we assume $C_{DH} = C_{DH} / C_{HD}$ and $C_{HD} = 1$, as this will not change the final results. Typically, we require $C_{DH} > 1$ because the cost of misclassifying diseased patients as healthy is higher than that of misclassifying healthy subject as diseased.

In this section, we formally address the problem of cost-sensitive feature selection by explicitly incorporating different misclassification costs in the objective function of the feature selection algorithm. Specifically, we will extend a widely used filter-type feature selection algorithm, i.e., Variance Score (VS) [11], to the corresponding cost-sensitive version (CostVS). It is worth noting, however, similar idea can also be used to derive other cost-sensitive feature selection algorithms.

Let $f_{r,i}$ denote the r -th feature of the i -th example \mathbf{x}_i . Then, we can define the scoring function of CostVS as maximizing

$$CostVS_r = \frac{1}{2m^2} \sum_{i=1}^m \sum_{j=1}^m cost(i, j)(f_{r,i} - f_{r,j})^2 \quad (1)$$

Where $cost(i, j)$ denotes the cost of misclassifying the i -th example as the j -th example. Because we use the class-dependent cost in this paper, those values can be easily gotten from C_{DH} and C_{HD} . Specifically, we have

$$cost(i, j) = \begin{cases} C_{DH} & \text{if } y_i = 'D' \text{ and } y_j = 'H' \\ C_{HD} & \text{if } y_i = 'H' \text{ and } y_j = 'D' \\ 0 & \text{elsewhere} \end{cases} \quad (2)$$

Where y_i and y_j are the class labels (‘D’ or ‘H’) of the i -th and j -th examples, respectively. Intuitively, in Eq. 1, we want to increase the contribution from those examples with higher misclassification cost, i.e., diseased patients with AD or MCI, such that the features which are more related to AD or MCI diseases can be selected.

2.2 Multimodal Data Fusion

To effectively fuse data from MRI, PET and CSF modalities, we adopt a multiple-kernels combination scheme in this paper. Assume we have m training examples, and each example has M modalities of data (i.e., $M=3$ in this paper), represented as $\mathbf{x}_i=\{\mathbf{x}_i^{(1)}, \dots, \mathbf{x}_i^{(p)}, \dots, \mathbf{x}_i^{(M)}\}$, $i=1, \dots, m$. Let $k^{(p)}(\cdot, \cdot)$ denote the Mercer kernel function on the p -th modality, then we can define the kernel function $k(\cdot, \cdot)$ on two data \mathbf{x} and \mathbf{z} as

$$k(\mathbf{x}, \mathbf{z}) = \sum_{p=1}^M \beta_p k^{(p)}(\mathbf{x}^{(p)}, \mathbf{z}^{(p)}) \quad (3)$$

Where β_p s are the nonnegative weighting parameters used to balance the contributions of different modalities. All β_p s are constrained by $\sum_p \beta_p = 1$.

Once we have defined the kernel function $k(\cdot, \cdot)$ on multimodal data, the m by m kernel matrix \mathbf{K} on the training multimodal data can be straightforwardly obtained as $\mathbf{K}=\{k(\mathbf{x}_i, \mathbf{x}_j)\}$. Then, the subsequent classifier can be directly built with the kernel matrix \mathbf{K} .

2.3 Cost-sensitive Classification

In this paper, we adopt support vector machine (SVM) implemented in LibSVM [12] with a threshold-moving strategy [9-10] as the cost-sensitive classifier, denoted as CostSVM in the rest of paper. Without loss of generality, we assume in this paper that the diseased patients are in the positive class (+1) while the healthy subjects are in the negative class (-1). Threshold-moving strategy moves the output threshold toward inexpensive class such that examples with higher costs become harder to be misclassified. Specifically, in the case of SVM classification, a positive constant is added to the threshold of SVM decision function such that the function value increases towards the positive class. It is worth noting that the threshold-moving is a post-processing strategy which introduces cost-sensitivity at test stage.

3 Experiments

In this section, we evaluate the effectiveness of the proposed cost-sensitive feature selection method (CostVS) and multimodal classification method (MultiCost) on the Alzheimer's Disease Neuroimaging Initiative (ADNI) database.

3.1 Experimental Settings

The ADNI database (www.loni.ucla.edu/ADNI) contains approximately 200 cognitively normal elderly subjects to be followed for 3 years, 400 subjects with MCI to be followed for 3 years, and 200 subjects with early AD to be followed for 2 years. In this paper, we focus on multimodal classification of AD and MCI converters who convert to AD after some years. Accordingly, corresponding subjects with all MRI, PET, and CSF data at baseline are included. This yields a total of 146 subjects, including 51 AD patients, 43 MCI patients who had converted to AD within 18 months, and 52 healthy controls (HCs). Standard image pre-processing is performed

for all MRI and PET images, including anterior commissure (AC) - posterior commissure (PC) correction, skull-stripping, removal of cerebellum, and segmentation of structural MR images into three different tissues: grey matter (GM), white matter (WM), and cerebrospinal fluid (CSF). With atlas warping, we can partition each subject image into 93 regions of interests (ROIs) [13]. For each of the 93 ROIs, we compute the GM tissue volume from the subject's MRI image. For PET image, we first rigidly align it with its respective MRI image of the same subject, and then compute the average value of PET signals in each ROI. Therefore, for each subject, we can finally obtain totally 93 features from MRI image, other 93 features from PET image, and 3 features ($A\beta_{42}$, t-tau, and p-tau) from CSF biomarkers.

To evaluate the performances of different algorithms, we use 10-fold cross-validation strategy to compute *the total cost* (cost) in misclassifying subjects, *the classification accuracy* (ACC) for measuring the proportion of subjects correctly classified among the whole population, *the sensitivity* (SEN) for measuring the proportion of AD or MCI patients correctly classified, and *the specificity* (SPE) for measuring the proportion of healthy subjects correctly classified. To compute the total cost, we use the fixed costs of $C_{DH} = 20$ and $C_{HD} = 1$.

In our experiments, we compare four algorithms for multimodal classification, which are built based on a similar flowchart as MultiCost; and the differences lie in that they use different feature selection and classification methods:

- **VS_SVM** -- VS feature selection plus SVM;
- **VS_CostSVM** -- VS feature selection plus cost-sensitive SVM (CostSVM);
- **CostVS_SVM** -- CostVS feature selection plus SVM;
- **MultiCost** (CostVS_CostSVM) -- CostVS feature selection plus CostSVM.

All algorithms use a linear kernel to compute the kernels, and the values of the weighting parameters β_m s are gotten through cross-validation on training data using the grid search. For VS_CostSVM and MultiCost, the optimal values of the positive constant are tuned on the training data to achieve the highest sensitivity while maintaining similar accuracy as the corresponding cost-blind classifier.

3.2 Results

Fig. 1 shows the performances (cost, accuracy, and sensitivity) of the four algorithms under different number of features selected for AD classification. As can be seen from Fig. 1, compared with VS_SVM and VS_CostSVM, both MultiCost and CostVS_SVM greatly reduce the cost and at the same time improve the accuracy, as well as sensitivity and specificity, especially when a small number of selected features is used. Fig. 1 also indicates that by using the threshold-moving strategy, MultiCost further improves the sensitivity while keeping a similar accuracy as CostVS_SVM.

To make a further comparison among the four algorithms, we averaged the performances under different numbers of selected features (from 10 features to all features) and the results are given in Table 1. Here, we do not consider the cases of using less than 10 features, because all algorithms achieve bad performance. Moreover, we perform significance test between MultiCost (or CostVS_SVM) and other method on all performance measures, for both AD and MCI classifications. It can be seen from Table 1 that MultiCost always achieves significantly better results

than other methods on cost and sensitivity measures, while keeping accuracy similar to CostVS_SVM but better than VS_SVM and VS_CostSVM. On the other hand, Table 1 shows that both MultiCost and CostVS_SVM have much lower deviation than VS_SVM and VS_CostSVM, and thus are more robust to the variations on the number of selected features.

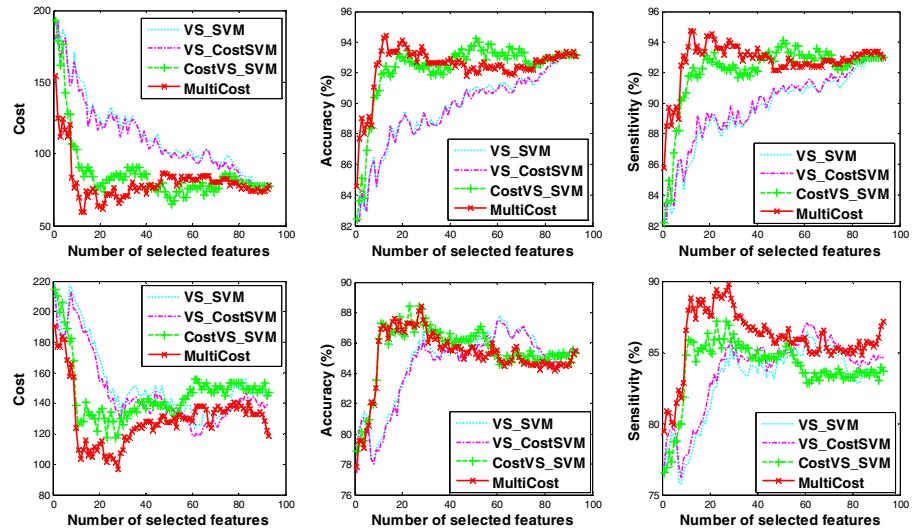


Fig. 1. Comparison of performance of four different methods on multimodal-data based AD (top) and MCI (bottom) classification

Table 1. Comparison of the averaged performance of different algorithms on multimodal data. The values in brackets are the standard deviation. The symbol * (or $\sqrt{ }$) denotes that the difference between MultiCost (or CostVS_SVM) and other method is significant (at 95% significance level using paired t-test). The best value in each column is bolded.

Methods	AD				MCI			
	cost	ACC (%)	SEN (%)	SPE (%)	cost	ACC (%)	SEN (%)	SPE (%)
VS_SVM	107.2 (19.0)	90.4 (1.8)	90.2 (1.8)	90.5 (1.8)	148.5 (17.9)	85.2 (1.8)	83.6 (2.0)	86.5 (1.6)
VS_CostSVM	104.3 (18.1)	90.3 (1.8)	90.5 (1.7)	90.0 (1.9)	142.2 (17.7)	85.0 (1.8)	84.4 (2.0)	85.6 (1.7)
CostVS_SVM	80.1 (6.4)	92.9 (0.6)	92.7 (0.6)	93.0 (0.6)	140.9 (9.6)	86.0 (1.0)	84.5 (1.1)	87.3 (0.9)
MultiCost	76.8 (6.3)	92.8 (0.6)	93.1 (0.6)	92.5 (0.6)	124.9 (11.7)	85.7 (1.1)	86.5 (1.3)	84.9 (0.9)

Fig. 1 and Table 1 show that VS_CostSVM achieves a similar performance as VS_SVM. Under the constraint of keeping a similar accuracy as VS_SVM, VS_CostSVM slightly improves the sensitivity at the price of a slight reduction in specificity, resulting in a slight improvement on the total cost. This implies that VS_CostSVM alone which incorporates the cost information only at the classification stage is not sufficient, due to the complexity and high-dimensionality of brain imaging data. In contrast, the MultiCost method exploits the cost information at both feature selection and classification stages, and thus significantly improves the performance.

Finally, we test the capability of the proposed CostVS feature selection method in selecting AD-related features, compared with standard VS method. In a typical experiment, we find, among its top 12 selected features, VS can only detect 4 AD-related features, computed from the ROIs such as ‘supramarginal gyrus right’, ‘entorhinal cortex right’, ‘cingulate right’, and ‘temporal pole left’, while CostVS can detect 7 AD-related features, computed from ROIs such as ‘entorhinal cortex right’, ‘hippocampal formation right’, ‘amygdala right’, ‘parahippocampal gyrus right’, ‘parahippocampal gyrus left’, ‘hippocampal formation left’, and ‘amygdala left’. This shows that CostVS has more power than VS in detecting AD-related brain regions for guiding AD classification. Fig. 2 plots the top 12 features (brain regions) detected by CostVS.

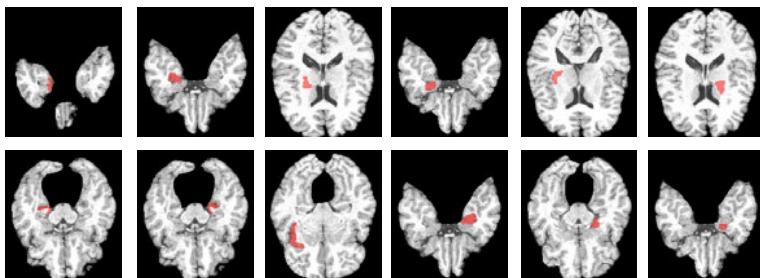


Fig. 2. Top 12 regions (highlighted) detected by CostVS. Each region is shown in a different view to enhance the visual quality.

4 Conclusion

This paper addresses the problem of exploiting the different misclassification cost in AD or MCI classification. Although using cost information for aiding classification is common in other areas such as machine learning and data mining, to our best knowledge, this issue is rarely investigated in AD-related studies. To effectively use the cost information for multimodal AD or MCI classification, in this paper we propose a Multi-stage Cost-sensitive classification method (MultiCost) to exploit cost information at both feature selection and classification stages. Experimental results on ADNI dataset validate the efficacy of the proposed method. In future work, we will extend our cost-sensitive feature selection to other feature selection methods and will also test other cost-sensitive classifiers for further improvement of classification performance.

Acknowledgments. This work was supported in part by NIH grants EB006733, EB008374, EB009634 and MH088520, and also by National Science Foundation of China under grant No. 60875030.

References

1. Ron, B., Elizabeth, J., Kathryn, Z.G., Arrighi, H.M.: Forecasting the global burden of Alzheimer's disease. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association* 3, 186–191 (2007)
2. Walhovd, K.B., Fjell, A.M., Dale, A.M., McEvoy, L.K., Brewer, J., Karow, D.S., Salmon, D.P., Fennema-Notestine, C.: Multi-modal imaging predicts memory performance in normal aging and cognitive decline. *Neurobiol. Aging* 31, 1107–1121 (2010)
3. Vemuri, P., Wiste, H.J., Weigand, S.D., Shaw, L.M., Trojanowski, J.Q., Weiner, M.W., Knopman, D.S., Petersen, R.C., Jack Jr., C.R.: MRI and CSF biomarkers in normal, MCI, and AD subjects: predicting future clinical change. *Neurology* 73, 294–301 (2009)
4. Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehericy, S., Habert, M.O., Chupin, M., Benali, H., Colliot, O.: Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *Neuroimage* (2010), doi:10.1016/j.neuroimage.2010.06.013
5. Kloppel, S., Stonnington, C.M., Chu, C., Draganski, B., Scahill, R.I., Rohrer, J.D., Fox, N.C., Jack Jr., C.R., Ashburner, J., Frackowiak, R.S.: Automatic classification of MR scans in Alzheimer's disease. *Brain* 131, 681–689 (2008)
6. Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C.: COMPARE: Classification Of Morphological Patterns using Adaptive Regional Elements. *IEEE Trans. Medical Imaging* 26, 93–105 (2007)
7. Hinrichs, C., Singh, V., Xu, G., Johnson, S.: MKL for robust multi-modality AD classification. In: Yang, G.-Z., Hawkes, D., Rueckert, D., Noble, A., Taylor, C. (eds.) MICCAI 2009. LNCS, vol. 5762, pp. 786–794. Springer, Heidelberg (2009)
8. Davatzikos, C., Bhatt, P., Shaw, L.M., Batmanghelich, K.N., Trojanowski, J.Q.: Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiol. Aging* (2010) (in press)
9. Elkan, C.: The foundations of cost-sensitive learning. In: The 17th International Joint Conference on Artificial Intelligence, pp. 973–978 (2001)
10. Zhou, Z.H., Liu, X.Y.: Training cost-sensitive neural networks with methods addressing the class imbalance problem. *IEEE Trans. Knowledge and Data Engineering* 18, 63–77 (2006)
11. Zhang, D., Chen, S., Zhou, Z.-H.: Constraint Score: A new filter method for feature selection with pairwise constraints. *Pattern Recognition* 41(5), 1440–1451 (2008)
12. Chang, C.C., Lin, C.J.: LIBSVM: a library for support vector machines (2001)
13. Zhang, D., Wang, Y., Zhou, L., Yuan, H., Shen, D.: Multimodal classification of Alzheimer's disease and mild cognitive impairment. *NeuroImage* 55(3), 856–867 (2011)

Classifying Small Lesions on Breast MRI through Dynamic Enhancement Pattern Characterization

Mahesh B. Nagarajan^{1,*}, Markus B. Huber¹, Thomas Schlossbauer²,
Gerda Leinsinger², Andrzej Krol³, and Axel Wismüller¹

¹ Departments of Imaging Sciences and Biomedical Engineering, University of Rochester, Rochester NY 14627, USA
nagaraja@bme.rochester.edu

² Department of Radiology, Ludwig Maximilians Universität,
80336 München, Germany

³ Department of Radiology, SUNY Upstate Medical University,
Syracuse NY 13210, USA

Abstract. Dynamic characterization of the lesion enhancement pattern can improve the classification performance of small diagnostically challenging lesions on dynamic-contrast enhanced MRI. This involves extraction of texture features from all post-contrast images of the lesion rather than using the first post-contrast image alone. In this study, statistical texture features derived from gray-level co-occurrence matrices are extracted from all five post-contrast images of 60 lesions and then used in a supervised learning task with a support vector regressor. Our results show that this approach significantly improves the performance of classifying small lesions ($p < 0.05$). This suggests that such dynamic characterization of lesion enhancement has significant potential in assisting breast cancer diagnosis for small lesions.

Keywords: dynamic breast MRI, texture analysis, dynamic enhancement characterization, gray-level co-occurrence matrix, support vector regression.

1 Introduction

Dynamic Contrast-Enhanced Magnetic Resonance Imaging (DCE-MRI) is an important tool in breast cancer diagnosis. The ability of CADx applications to achieve a high diagnostic sensitivity (upto 97%) and reasonable specificity (76.5%) in the task of classifying suspicious lesions on DCE-MRI through dynamic or morphological criteria is well established[1]. However, not many studies have focused on evaluating the value of DCE-MRI in small lesions where lesions may not exhibit typical characteristics of benign and malignant tumors[2]. In this regard, Leinsinger et al. 2006 reported a diagnostic accuracy of 75% in detecting breast cancer through cluster analysis of signal intensity time curves [3]. More

* Corresponding author.

recently, Schlossbauer et al. 2008 reported an AUC of 0.76 when using dynamic characteristics extracted from a dataset of small lesions (mean size 1.1 cm) in classifying lesion character, and an AUC of 0.61 when using morphologic criteria for the character classification task [2]. The primary goal of this work was to improve the classification performance of such small, diagnostically challenging lesions on DCE-MRI.

Texture analysis can be used to quantify image patterns from a specified region of interest (ROI) [4]. This study uses second-order statistical texture features derived from gray-level co-occurrence matrices (GLCM) as described in the seminal work of Haralick [5,6]. GLCM has found use in a wide scope of medical image analysis tasks such as distinguishing pathological patterns from healthy lung tissue on chest CT [7], classifying healthy and abnormal tissue on radiographic mammography images [8] and quantitative analysis of carotid atherosclerotic plaques on ultrasound B-mode scans [9]. Prior studies have shown that GLCM-derived texture features can extract valuable information from the lesion and can contribute to high diagnostic accuracy in the task of classifying lesion character [10,11,12].

However, texture analysis in such studies has focused on extracting texture features from a single post-contrast image (usually the first of five). This approach ignores textural information that could be obtained from observing lesions over time, specifically from later post-contrast images of the lesion. We hypothesize that providing the classifier with supplementary information derived through performing texture analysis on later post-contrast images can significantly improve the performance of lesion character classification, specifically of small, diagnostically challenging lesions, as used in this study.

In this work, texture analysis using GLCM is performed on all five post-contrast images of a dynamic breast MRI exam and the texture features extracted are combined to form lesion characterizing 5-D texture feature vectors. Such a *dynamic characterization of the lesion enhancement pattern* is subsequently followed by a classification task which employs a support vector regressor (SVR)[13]. The purpose of this study was to evaluate whether dynamic characterization of the lesion enhancement pattern would significantly improve the classification performance over using the texture features from the first post-contrast image alone.

2 Data

Sixty lesions were identified from a representative set of dynamic contrast-enhanced breast MRI exams from 54 female patients. The mean patient age was 52 with a standard deviation of 12 and a range of 27 to 78. In all cases, histo-pathologically confirmed diagnosis from needle aspiration/excision biopsy was available prior to this study; 32 of the lesions were diagnosed as benign and the remaining 28 as malignant. Mean lesion diameter was 1.05 cm (standard deviation of 0.73 cm).

Patients were scanned in the prone position using a 1.5T MR system (Magnetom VisionTM, Siemens, Erlangen, Germany) with a dedicated surface coil

to enable simultaneous imaging of both breasts. Images were acquired in the transversal slice orientation using a T1-weighted 3D spoiled gradient echo sequence with the following imaging parameters; echo repetition time (TR) = 9.1 ms, echo time (TE) = 4.76 ms and flip angle (FA) = 25°. Acquisition of the pre-contrast series was followed by the administration of 0.1 mmol/kg body weight of paramagnetic contrast agent (gadopentate dimeglumine, MagnevistTM, Schering, Berlin, Germany). Five post-contrast series were then acquired, each with a measurement time of 83 seconds, and at intervals of 110 seconds. All studies were acquired with informed consent from the patients and were evaluated in this study in a retrospective manner.

In the collection of patient data used in this study, images in the dynamic series were acquired with two different settings of spatial parameters; for 19 patients, the images were acquired as 32 slices per series with a 512 x 512 matrix, 0.684 x 0.684 mm² in-plane resolution and 4 mm slice thickness, while in other cases, the same images were acquired as 64 slices per series with a 256 x 256 matrix, 1.37 x 1.37 mm² and 2 mm slice thickness. To maintain uniform image data for texture analysis, the images acquired with a 512x512 matrix were reduced to a 256x256 matrix through bilinear interpolation.

3 Methods

3.1 Lesion Annotation and Pre-processing

All lesions were localized manually by a radiologist on subtraction images created by subtracting the pre-contrast image from the fourth post-contrast image. With the exception of two patients, where two separate lesions were chosen for analysis, one primary lesion was selected from all patients for analysis. In four cases, these lesions were captured with two non-overlapping ROIs; a single encapsulating ROI was used to capture the lesion in the rest. Each identified lesion was extracted as a 2-D square region of interest (ROI) of dimensions 11x11 pixels on the central slice of the lesion from the pre-contrast and the all post-contrast images of the T1 dynamic series. Three examples of such lesions are shown in Figure 1. A fuzzy C-Means (FCM) approach was then used to segment the lesions in the square ROI[14].

Lesion enhancement normalization was performed on each post-contrast ROI (S_i) by subtracting and dividing the i^{th} post-contrast ROI S_i $i = \{1, 2, 3, 4, 5\}$, with the corresponding ROI annotated on the pre-contrast lesion (S_0), i.e. $(S_i - S_0)/S_0$. Following normalization, the ROIs were re-binned to 32 gray-level histogram bins between 'global' minimum and maximum intensity limits as recommended by previous work[11].

3.2 Texture Analysis

Gray-level co-occurrence matrices (GLCM) were extracted from the lesion ROIs as described in [5]. An inter-pixel distance of $d = 1$ was used in generating the GLCMs to maximize counting statistics. On each ROI, GLCMs were generated

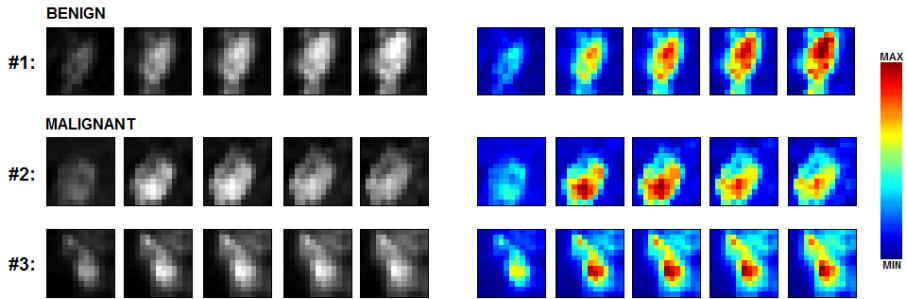


Fig. 1. Examples each of benign (1) and malignant (2, 3) lesion ROIs at five different stages of contrast uptake. The gray-level images were used in texture analysis after lesion segmentation; the corresponding color images better showcase the dynamic contrast uptake signatures of these lesions (persistent increase (1) for benign; washout (2) and plateau (3) patterns for malignant). The red pixels correspond to high intensity values and blue to low values.

in the four directions mentioned earlier and then summed up element-wise resulting in one non-directional GLCM from which the Haralick features f1-f13 were computed. Each texture feature was computed on every post-contrast image and then combined into a texture feature vector; 13 such texture feature vectors with a dimensionality of 5 were computed in this manner for every individual lesion ROI. For comparison, these 13 texture features were also extracted from the first post-contrast image alone.

3.3 Classification

The extraction of texture features and subsequent feature selection was followed by a supervised learning step where the lesion patterns were classified as benign or malignant. In this work, a support vector regressor with a radial basis function kernel (SVRrbf) and a support vector regressor with a linear kernel (SVRlin) were used. A fuzzy k-Nearest Neighbor (k-NN) classifier was used as a baseline for comparison with support vector regression. 70% of the data was randomly sub-sampled for the training phase while the remaining 30% served as an independent test set. The training data was sub-sampled from the complete dataset in such a manner that at least 40% of each class was represented. Special care was taken to ensure that lesion ROIs extracted from the same patient were used either as training or test data to prevent any potential for biased training. To ensure the integrity of the independent test set, global intensity limits for pre-processing were determined using lesion ROIs from the training data alone. In the training phase, models were created from labeled data by employing a random sub-sampling cross-validation strategy to optimize the free parameters for the classifiers (the number of nearest neighbors 'k' for fkNN, the cost parameter for SVRlin and SVRrbf, and the shape parameter of the radial basis function kernel

of SVRrbf) while creating the best model that separates both classes. Then, during the testing phase, the optimized classifier predicted the label (benign or malignant) of lesion ROIs in the independent test dataset; an ROC curve was generated and used to compute the area under the curve (AUC) which served as a measure of classifier performance. This process was repeated 100 times resulting in an AUC distribution for each feature set.

A Wilcoxon signed-rank test was used to compare two AUC distributions of the same feature under different experimental conditions while a Mann-Whitney U-test was used to compare different features. Significance thresholds were adjusted for multiple comparisons using the Holm-Bonferroni correction. Texture, classifier and statistical analysis were implemented using Matlab 2008b (The MathWorks, Natick, MA).

4 Results

Figure 2 shows the classification performance obtained using three different classifiers - fkNN, SVRlin and SVRrbf, for texture features f4 and f6 which exhibited the best overall AUC values (0.82). The best classification performance was obtained with the SVRrbf classifier when these texture features were extracted from all five post-contrast images. When texture features were extracted from the first post-contrast image alone, the performance of fkNN was marginally comparable to that of SVRrbf.

Table 1 compares the classification performance obtained with SVRrbf for texture features extracted from all five post-contrast images as opposed to that obtained when texture analysis involves the first post-contrast image alone. Six of thirteen texture features show significant improvements in classifier performance when the dynamic characterization approach was used. In particular, texture features f4 and f6 exhibited an AUC of 0.82 which was the highest AUC value observed in this study. Only one texture feature f2 significantly deteriorated in performance when the dynamic texture quantification approach was pursued.

5 Discussion

The primary goal of this study was to improve the performance of classifying diagnostically challenging lesions, specifically those considered small (mean lesion diameter of 1.05 cm), on DCE-MRI. Previous approaches to CADx which involved quantifying the lesion enhancement pattern of breast lesions on the first post-contrast image alone using texture features [10,11,12] do not provide the best classification performance in this context (as confirmed in this study). To address this problem, we introduce a method to dynamically characterize the lesion enhancement pattern by extracting texture features from all five post-contrast images of the lesion. Our results show that such an approach can significantly improve the performance of the lesion character classification task.

The improvement in classification performance with the dynamic characterization approach (table 1) is attributed to more lesion pattern-specific information

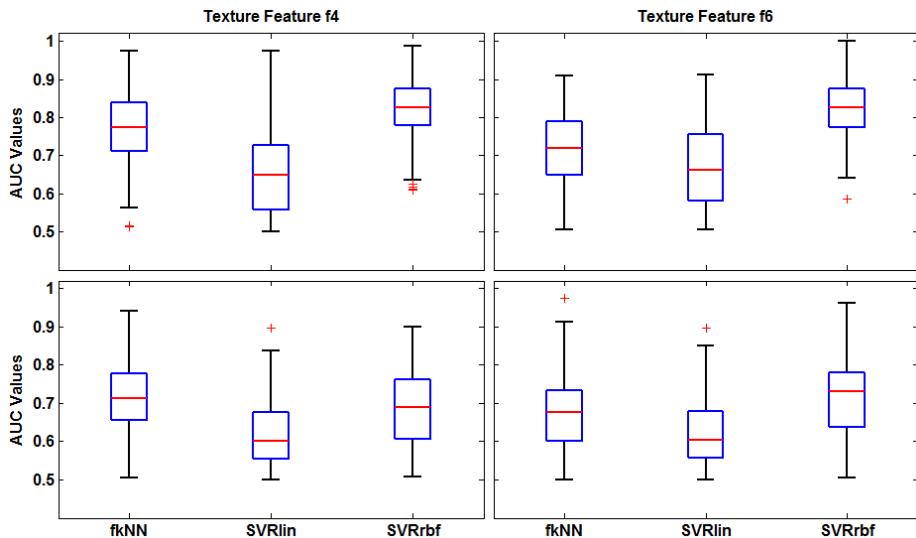


Fig. 2. Classification performance of texture features f4 (left) and f6 (right), as observed for different classifiers when all five post-contrast images are used for the classification task. The top row corresponds to texture feature vectors extracted using all five post-contrast image (ALL) while the bottom row corresponds to texture features extracted from the first post-contrast image alone (P1). For each distribution, the central mark corresponds with the median and the edges are the 25th and 75th percentile.

being incorporated into the lesion character classification task. Such 5-D texture feature vectors not only quantify lesion texture patterns but also how these properties vary with contrast uptake dynamics (Figure 1). We believe this to be primary reason behind the improvements observed in lesion character classification. The SVRrbf classifier yielded the best results when the lesion enhancement pattern was characterized dynamically suggesting that it is able to better use the supplementary information provided by texture features extracted from later post-contrast images in distinguishing between benign and malignant lesions.

This work revealed certain limitations of using Haralick texture features for characterizing the enhancement pattern in small lesions on DCE-MRI. While Haralick describes 14 texture features in his work [5], the feature f14 (Maximal Correlation Coefficient) was undefined for the lesions used in this study owing to the small size of lesions. The recommended pre-processing steps also resulted in a constant image for these ROIs on certain post-contrast images which yields undefined values for texture features f3 (correlation) and f12 (information measure of correlation I). This analysis excluded such lesions from the analysis for these features (results marked with a star). As an outlook for future research, one could replace such statistical features with more sophisticated topological or geometrical texture features.

Table 1. Classification performance (mean AUC \pm standard deviation) of texture features extracted from the 1st post-contrast image (P1) is compared to that of texture feature vectors created from all five post-contrast images (ALL). Significantly higher AUC values in each row are marked in bold. Results marked with an 'x' indicate numbers that were obtained after excluding two lesions for reasons mentioned in the text.

Features	P1	ALL	p	threshold
f1	0.66 ± 0.09	0.66 ± 0.09	0.2868	—
f2	0.68 ± 0.12	0.63 ± 0.09	0.0002	0.0071
f3	0.60 ± 0.08	0.60 ± 0.08^x	0.4172	—
f4	0.69 ± 0.10	0.82 ± 0.08	< 0.0001	0.0042
f5	0.69 ± 0.10	0.69 ± 0.10	0.4006	—
f6	0.72 ± 0.10	0.82 ± 0.08	< 0.0001	0.0038
f7	0.67 ± 0.10	0.79 ± 0.09	< 0.0001	0.0045
f8	0.61 ± 0.08	0.72 ± 0.11	< 0.0001	0.0050
f9	0.68 ± 0.09	0.72 ± 0.11	< 0.0001	0.0063
f10	0.67 ± 0.11	0.64 ± 0.10	0.0345	—
f11	0.73 ± 0.10	0.74 ± 0.08	0.6182	—
f12	0.62 ± 0.08	0.64 ± 0.09^x	0.3120	—
f13	0.60 ± 0.07	0.67 ± 0.10	< 0.0001	0.0056

Another limitation of this study regards the use of 2D lesion ROIs in texture analysis owing to the anisotropy of the pixels involved. While previous research has shown that volumetric analysis of lesions improves classification performance[11], arguments have been made against acquiring breast images with isotropic voxels owing to the longer imaging time involved as well as the smaller coverage of the area being imaged[12]. Future studies could also use feature selection techniques to explore the contribution of each post-contrast image to the overall classification performance.

In conclusion, this study shows that the performance of automated character classification of diagnostically challenging lesions, specifically those considered small (mean lesion diameter of 1.05 cm), can be significantly improved through dynamic characterization of the lesion enhancement pattern.

Acknowledgments. This research was funded in part by the Clinical and Translational Science Award 5-28527 within the Upstate New York Translational Research Network (UNYTRN) of the Clinical and Translational Science Institute (CTSI), University of Rochester, and by the Center for Emerging and Innovative Sciences (CEIS), a NYSTAR-designated Center for Advanced Technology.

References

1. Fischer, D., Wurdinger, S., Boettcher, J., Malich, A., Kaiser, W.: Further signs in the evaluation of magnetic resonance mammography: a retrospective study. *Investigative Radiology* 40(7), 430–435 (2005)
2. Schlossbauer, T., Leinsinger, G., Wismüller, A., Lange, O., Scherr, M., Meyer-Baese, A., Reiser, M.: Classification of small contrast enhancing breast lesions in dynamic magnetic resonance imaging using a combination of morphological criteria and dynamic analysis based on unsupervised vector-quantization. *Investigative Radiology* 43(1), 56–64 (2008)
3. Leinsinger, G., Schlossbauer, T., Scherr, M., Lange, O., Reiser, M., Wismüller, A.: Cluster analysis of signal-intensity time course in dynamic breast MRI: does unsupervised vector quantization help to evaluate small mammographic lesions? *European Radiology* 16(5), 1138–1146 (2006)
4. Lerski, R.A., Straughan, K., Schad, L.R., Boyce, D., Blüml, S., Zuna, I.: Tissue characterization by magnetic-resonance spectroscopy and imaging - results of a concerted research-project of the european-economic-community.8. MR image texture analysis - an approach to tissue characterization. *Magnetic Resonance Imaging* 11(6), 873–887 (1993)
5. Haralick, R.M., Shanmuga, K., Dinstein, I.: Textural features for image classification. *IEEE Transactions on Systems Man and Cybernetics Smc3(6)*, 610–621 (1973)
6. Haralick, R.M.: Statistical and structural approaches to texture. *Proceedings of the IEEE* 67(5), 786–804 (1979)
7. Korfiatis, P., Kalogeropoulou, C., Karahaliou, A., Kazantzis, A., Skiadopoulos, S., Costaridou, L.: Texture classification-based segmentation of lung affected by interstitial pneumonia in high-resolution CT. *Medical Physics* 35(12), 5290–5302 (2008)
8. Chan, H.P., Wei, D.T., Helvie, M.A., Sahiner, B., Adler, D.D., Goodsitt, M.M., Petrick, N.: Computer-aided classification of mammographic masses and normal tissue - linear discriminant-analysis in texture feature space. *Physics in Medicine and Biology* 40(5), 857–876 (1995)
9. Wilhjelm, J.E., Gronholdt, M.L.M., Wiebe, B., Jespersen, S.K., Hansen, L.K., Sillesen, H.: Quantitative analysis of ultrasound B-mode images of carotid atherosclerotic plaque: Correlation with visual classification and histological examination. *IEEE Transactions on Medical Imaging* 17(6), 910–922 (1998)
10. Gibbs, P., Turnbull, L.W.: Textural analysis of contrast-enhanced MR images of the breast. *Magnetic Resonance in Medicine* 50(1), 92–98 (2003)
11. Chen, W., Giger, M.L., Li, H., Bick, U., Newstead, G.M.: Volumetric texture analysis of breast lesions on contrast-enhanced magnetic resonance images. *Magnetic Resonance in Medicine* 58(3), 562–571 (2007)
12. Nie, K., Chen, J.H., Yu, H.J., Chu, Y., Nalcioglu, O., Su, M.Y.: Quantitative analysis of lesion morphology and texture features for diagnostic prediction in breast MRI. *Academic Radiology* 15(12), 1513–1525 (2008)
13. Drucker, H., Burges, C., Kaufman, L., Smola, A., Vapnik, V.: Support vector regression machines. In: *Advances in Neural Information Processing Systems*, vol. 9, pp. 155–161 (1996)
14. Chen, W.J., Giger, M.L., Bick, U.: A fuzzy c-means (FCM)-based approach for computerized segmentation of breast lesions in dynamic contrast-enhanced MR images. *Academic Radiology* 13(1), 63–72 (2006)

Computer-Aided Detection of Polyps in CT Colonography with Pixel-Based Machine Learning Techniques

Jian-Wu Xu and Kenji Suzuki

Department of Radiology, The University of Chicago
5841 South Maryland Avenue, Chicago, IL 60637
{jwxu,suzuki}@uchicago.edu

Abstract. Pixel/voxel-based machine-learning techniques have been developed for classification between polyp regions of interest (ROIs) and non-polyp ROIs in computer-aided detection (CADe) of polyps in CT colonography (CTC). Although 2D/3D ROIs can be high-dimensional, they may reside in a lower dimensional manifold. We investigated the manifold structure of 2D CTC ROIs by use of the Laplacian eigenmaps technique. We compared a support vector machine (SVM) classifier with the Laplacian eigenmaps-based dimensionality-reduced ROIs with massive-training support vector regression (MTSVR) in reduction of false positive (FP) detections. The Laplacian eigenmaps-based SVM classifier removed 16.0% (78/489) of FPs without any loss of polyps in a leave-one-lesion-out cross-validation test, whereas the MTSVR removed 49.9% (244/489); thus, yielded a 96.6% by-polyp sensitivity at an FP rate of 2.4 (254/106) per patient.

Keywords: colorectal cancer, computer-aided detection, manifold learning, pixel-based machine learning, support vector machines.

1 Introduction

Computer-aided detection (CADe) of lesions in medical images is an active research area in medical imaging. One of the major challenges in current CADe of polyps in CT colonography (CTC) is to improve the specificity without sacrificing the sensitivity. If a large number of false-positive (FP) detections are produced by a CADe scheme, radiologists might lose their confidence in CADe. A CADe scheme [1] generally employs a classifier based on features extracted from each segmented candidate to distinguish between polyps and non-polyps. However, the extracted features might be noisy (with errors) due to CTC image reconstruction error, segmentation errors, and other factors. Moreover, it requires not only domain knowledge to design the set of features to be extracted, but also advanced feature selection methods for choosing the most discriminative ones. Recently, a pixel/voxel-based machine-learning technique based on nonlinear regression models has been

developed [2-4]. Although 2D/3D regions of interest (ROIs) are high-dimensional, they reside in a lower dimensional manifold because the variations of polyps and non-polyps are much smaller compared to the pixel/voxel dimension.

Because of the complexity of the subspace spanned by polyps and non-polyps, it is likely that the manifold is nonlinear. Recently, there have been many nonlinear dimensionality reduction techniques proposed in the literature, such as ISOMAP [5], locally linear embedding [6], Laplacian eigenmaps [7], and others. These techniques are able to represent the original high dimensional data by use of the nonlinear manifold in a much lower dimensional subspace. In this paper, we apply the Laplacian eigenmaps method to learn the manifold structure of the CTC ROIs. The manifold is learned based on the pairwise image similarity in ROIs whose locations are provided by the initial CADe scheme [8]. After the manifold learning step, we apply the support vector machine (SVM) classifier on the manifold space to classify between polyps and non-polyps in order to further reduce FP detections.

2 Methods

2.1 Manifold Learning

Given a set of high dimensional data points I_1, \dots, I_N in \Re^n , Laplacian eigenmaps aim at finding a set of data points x_1, \dots, x_N in a lower dimensional manifold \Re^d ($d < n$) such that x_i “represents” I_i well. We construct an adjacency graph G with N nodes. The i^{th} node corresponds to the i^{th} high dimensional data point. We connect nodes i to j with an edge if the Euclidean distance between the two nodes falls into the k nearest neighborhood. That is, an edge is put only for close k -neighbors. We set the weights to 1 for the connected edges and obtain the weight matrix W . From the weight matrix W , we can construct the graph Laplacian $L = D - W$ where D is a diagonal matrix with diagonal element as $D_{ii} = \sum_j W_{ij}$. The low-dimensional embedding x_i can be obtained by solving the generalized eigen-decomposition problem $LX = \lambda DX$.

2.2 Classification in Manifold Space

After the manifold structure is learned based on the Laplacian eigenmaps technique, we employ the SVM classifier to classify each ROI into polyps and non-polyps in order to further reduce FP detections from the initial CADe scheme. SVMs are a machine-learning technique that maximizes the margin of separation between positive and negative classes. An SVM as a memory-based learning method is very fast to train. As a memory-based learning method, the SVM classifier will retrain a set of training samples as “*support vectors*” after the training phase.

Given a set of N training data points $\{(x_i, y)\}_{i=1}^N$ where $x_i \in \Re^d$ is the vector in the learned manifold space corresponding to an individual ROI in the original CTC images, and $y_i \in \{-1, 1\}$ is the class label with 1 denoting polyps and -1 for non-polyps, the decision function for the SVM classifier can be written as

$$f(x) = \sum_{i=1}^N \alpha_i y_i K(x_i, x) + \alpha_0 \quad (1)$$

The parameters $\alpha_i \geq 0$ are called Lagrange multipliers that are optimized through quadratic programming. We use the Gaussian kernel function for $K(x_i, x_j)$.

The optimal Lagrange multipliers $\alpha_i \geq 0$ in the optimal decision boundary is computed through the maximization of the following objective function:

$$\max_{\alpha_i} \quad \sum_{i=1}^N \alpha_i - \frac{1}{2} \sum_{i=1}^N \sum_{j=1}^N \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (2)$$

subject to the following constraints:

$$\sum_{i=1}^N \alpha_i y_i = 0 \quad (3)$$

Where $\alpha_i \geq 0$ for $i = 1, 2, \dots, N$.

3 Materials

The database used in this study was acquired at the University of Chicago Medical Center. It consisted of 212 CTC datasets obtained from 106 patients. Each patient followed the standard CTC procedure with pre-colonoscopy cleansing and colon insufflations with room air or carbon dioxide. Both supine and prone positions were scanned. Each reconstructed CT section had a matrix size of 512×512 pixels, with an in-plane pixel size of 0.5–0.7 mm. Seventeen patients had 29 colonoscopy-confirmed polyps, 15 of which were 5–9 mm and 14 were 10–25 mm in size. The database was divided into a training set and a testing set. The training set consisted of 30 patients, 10 of which had 10 polyps. An initial CADe scheme for detection of polyps in CTC [8] was applied to the database. The CADe scheme is composed of 1) colon segmentation, 2) detection of polyp candidates based on the shape index and curvedness of the segmented colon, 3) calculation of 3D pattern features, and 4) classification of the polyp candidates as polyps or non-polyps based on quadratic discriminant analysis. The initial CADe scheme achieved a by-polyp sensitivity of 96.6% (28/29) with 4.6 (489/106) FPs per patient in the data set. The major sources of FPs included rectal tubes, stool, haustral folds, colonic walls, and the ileocecal valve. We extracted a 2D ROI on the axial plane in the CTC images centered in each location provided by the CADe scheme output. The centers corresponded to the geometrical centers of each detected candidates. The size of each 2D ROI is 32 by 32 pixels. The ROIs with 32 by 32 pixels were sufficiently large to cover all polyps in the database. We used 56 true-positive (TP) ROIs (corresponding to 28 polyps) and 489 FP ROIs as the data used in the study.

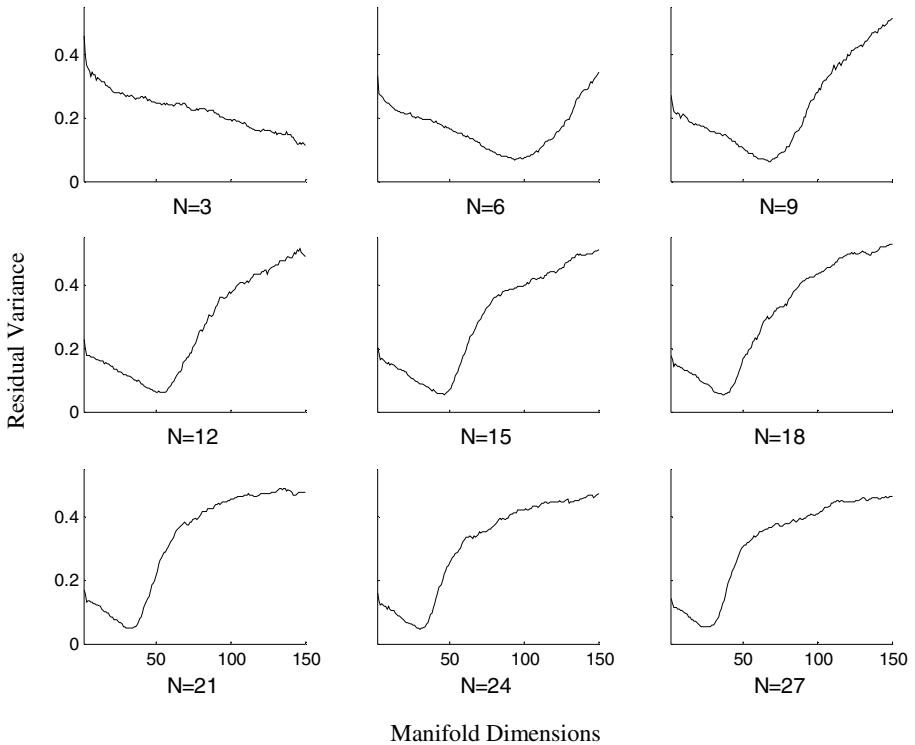


Fig. 1. Residual variances with respect to manifold space dimensions after applying the Laplacian eigenmaps with different neighborhood sizes (N) to the entire data set

4 Results

4.1 Optimal Nonlinear Embedding

We investigated the performance of Laplacian eigenmaps for dimension reduction from the pattern recognition perspective. To evaluate the optimal nonlinear embedding of Laplacian eigenmaps, we used the residual variance defined as

$$1 - R^2(\hat{L}, L) \quad (4)$$

where L is the Laplacian matrix of the original data in the high-dimensional space, \hat{L} is the matrix of the Euclidean distance in the low-dimensional embedding recovered by the algorithm, and R is the correlation coefficient over all entries of two matrices. The residual variance measures how well the low-dimensional embedding represents the original data in the high-dimensional space. Figure 1 plots the residual variance as a function of different numbers of dimensions after applying Laplacian eigenmaps with different numbers of neighborhood sizes to the entire data. Residual variance for neighborhood size 3 decreases as the dimensionality d increases. For neighborhood size greater than 3, the residual variance curves show “elbows” at

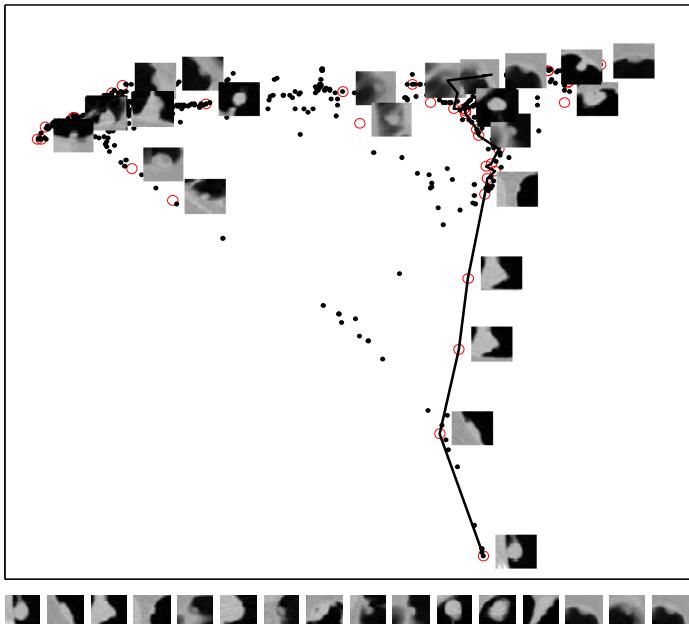


Fig. 2. Two dimensional Laplacian eigenmaps representation of CTC images with red circles denoting TPs and black dots standing for FPs. Sample TP ROIs are plotted. The ROIs below the figure show the smooth transition in appearance for those FPs along the manifold trajectory.

which the curves change from decreasing to increasing, which is an indication of intrinsic dimensionality. The minimal residual variance occurs when the neighborhood size is 24 and the manifold dimension is 30.

4.2 Manifold Visualization

We applied the Laplacian eigenmaps to the entire data set with the optimal parameters obtained in Section 3.1. Figures 2 and 3 plot the data in the first two dimensions in the optimal manifold with red circles denoting TP ROIs and black dots FP ROIs. Data points appear in a nice triangle shape with two clusters in the upper two vertices. Although there is no clear separation between two classes, TP ROIs are more concentrated around vertices, while FP ROIs are more spreading out. This might be due to the small TP ROIs available in the database. On the other hand, the small variation in the appearance of TP ROIs could also contribute to the distinct distribution. We also put some sample ROI images along the corresponding data points with TP images in Fig. 2 and FP images in Fig. 3. The images under the figures present the TPs and FPs along the two manifold trajectories in Figs. 2 and 3. The transitions in appearance in those ROIs can be observed.

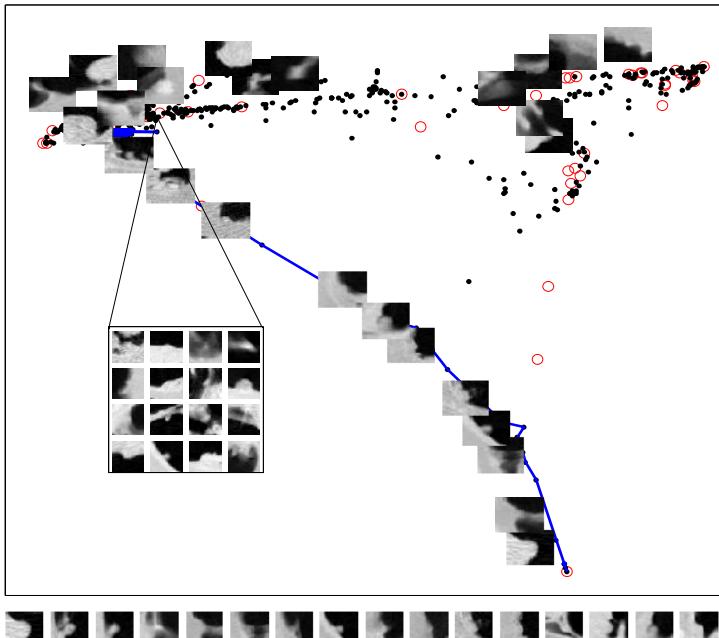


Fig. 3. Sample FP ROIs are plotted in the 2D manifold. The ROIs below the figure show the smooth transition in appearance for those FPs along the manifold trajectory.

4.3 Classification Performance

Although we can use the residual variance to determine the intrinsic dimensionality of the manifold for the 2D ROIs, optimal representation does not directly translate into optimal classification performance. Data visualization is a problem of unsupervised learning, while classification belongs to the supervised learning paradigm. Therefore, we varied the number of neighborhood sizes in the Laplacian eigenmaps and the manifold dimensions to decide the best parameters for the classification task. We used the area under the receiver-operating-characteristic curve (AUC) value as the performance metric. Figure 4 plots the AUC values versus different parameters. The maximum AUC value occurred when the neighborhood size in the Laplacian eigenmaps was 4 and the manifold dimension was 45 which were different from the parameters obtained from the optimal representation approach. Because we used the AUC criterion to learn the optimal parameter and dimensionality in a supervised approach, the parameters were different from the ones learned in the optimal embedding which was unsupervised learning.

We evaluated the overall classification performance in terms of FP reduction of the proposed method by use of the free-response receiver-operating-characteristic (FROC) analysis. Figure 5 presents the FROC curve for the SVM classification applied to the manifold learned through the Laplacian eigenmaps with the optimal parameters. By using the SVM classifier coupled with the Laplacian eigenmaps, 16.0% (78/489) of FPs were removed without any loss of polyps in a leave-one-lesion-out cross-validation test; thus,

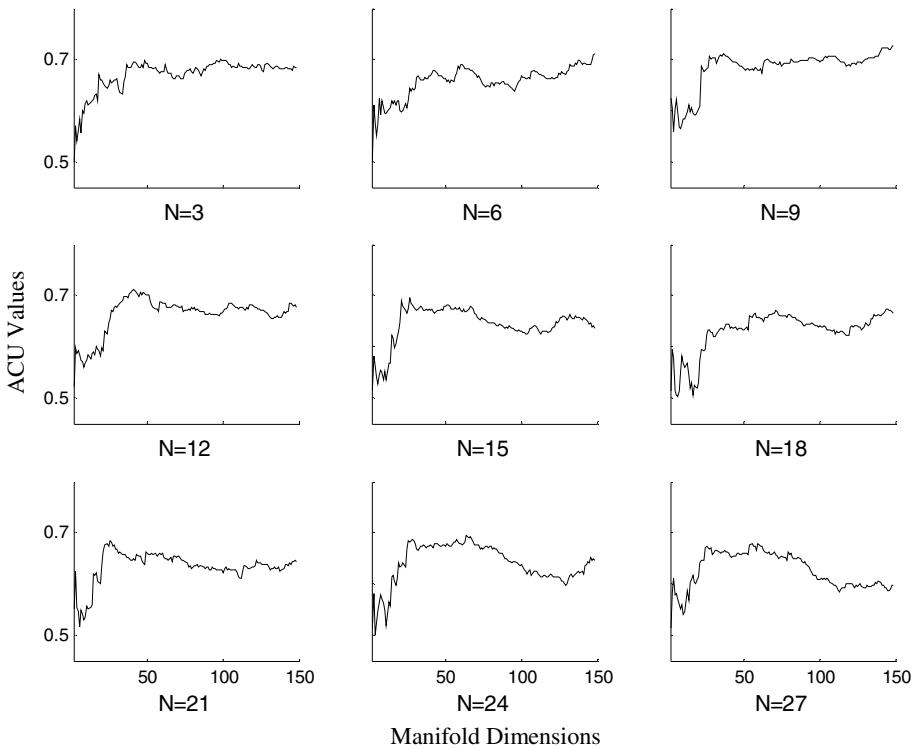


Fig. 4. AUC values obtained by the SVM classification applied to the manifold learned through the Laplacian eigenmaps with different neighborhood sizes (N)

the SVM classifier with Laplacian eigenmaps-based dimensionality-reduced ROIs achieved a 96.6% by-polyp sensitivity at an FP rate of 3.9 (411/106) per patient. We also compared the Laplacian eigenmaps against the principal component analysis (PCA). The PCA followed by SVM classification yielded a similar FP rate at the 96.6% sensitivity. However, the Laplacian eigenmaps-based SVM classifier obtained an AUC value of 0.80, while the PCA-based SVM classifier only had an AUC value of 0.67. The difference was statistically significant with a two-sided p value less than 0.05. Figure 5 also plots the FROC curve of MTSVR [3]. MTSVR achieved a performance of 2.4 FP per patient at the 96.6% by-polyp sensitivity. However, both MTSVR and the Laplacian eigenmaps-based SVM classifier yielded similar AUC values without any statistical significance.

5 Conclusion

We have investigated the manifold structure of the 2D ROIs in CTC images by use of the Laplacian eigenmaps technique. We compared SVM with the Laplacian eigenmaps-based dimensionality-reduced ROIs with the MTSVR in FP reduction. The MTSVR yielded a higher performance compared with other methods.

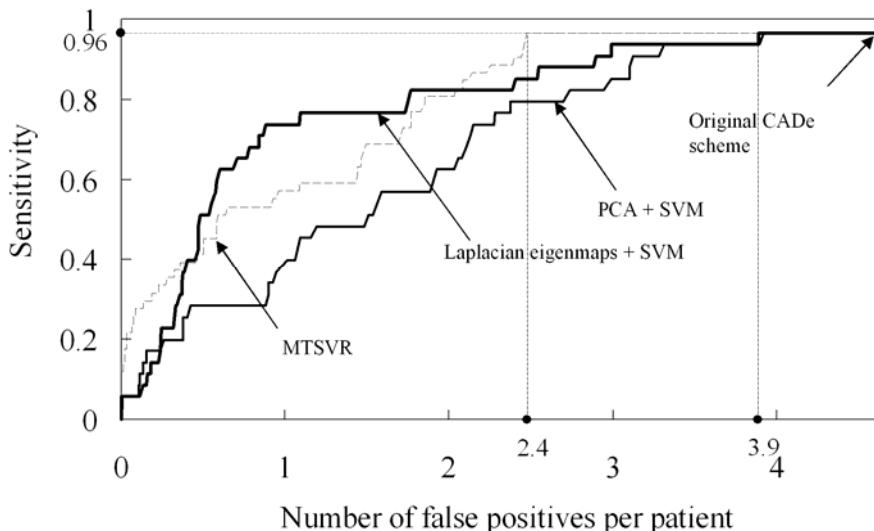


Fig. 5. FROC curves indicating the performance of the SVM classifier combined with the Laplacian eigenmaps, SVM classifier combined with PCA, and MTSVR.

Acknowledgments. This work was supported by Grant Number R01CA120549 from the NCI/NIH and partially by the NIH S10 RR021039 and P30 CA14599.

References

1. Summers, R.M., Beaulieu, C.F., Pusanik, L.M., Malley, J.D., Jeffrey Jr., R.B., Glazer, D.I., Napel, S.: Automated polyp detector for CT colonography: feasibility study. *Radiology* 216, 284–290 (2000)
2. Suzuki, K., Yoshida, H., Nappi, J., Dachman, A.H.: Massive-training artificial neural network (MTANN) for reduction of false positives in computer-aided detection of polyps: Suppression of rectal tubes. *Med. Phys.* 33, 3814–3824 (2006)
3. Xu, J., Suzuki, K.: Massive-Training Support Vector Regression and Gaussian Process for False-Positive Reduction in Computer-aided Detection of Polyps in CT Colonography. *Med. Phys.* 38, 1888–1902 (2011)
4. Suzuki, K., Zhang, J., Xu, J.: Massive-Training Artificial Neural Network Coupled with Laplacian-Eigenfunction-Based Dimensionality Reduction for Computer-Aided Detection of Polyps in CT Colonography. *IEEE Trans. on Med. Imaging* 29, 1907–1917 (2010)
5. Tenenbaum, J., de Silva, V., Langford, J.: A Global Geometric Framework for Nonlinear Dimensionality Reduction. *Science* 290, 2319–2323 (2000)
6. Roweis, S.T., Saul, L.K.: Nonlinear Dimensionality Reduction by Locally Linear Embedding. *Science* 290, 2323–2326 (2000)
7. Belkin, M., Niyogi, P.: Laplacian Eigenmaps for Dimensionality Reduction and Data Representation. *Neural Computation* 15, 1373–1396 (2003)
8. Yoshida, H., Nappi, J.: Three-dimensional computer-aided diagnosis scheme for detection of colonic polyps. *IEEE Trans. Med. Imaging* 20, 1261–1274 (2001)

Author Index

- Acosta, Oscar 142
Adluru, Nagesh 225
Aljabar, Paul 159
Anderson, Jeffrey S. 59
Arridge, Simon 167
Atkins, M. Stella 291
Awate, Suyash P. 59
Banerjee, Deb 265
Basso, Curzio 67
Benali, Habib 201
Bhole, Chetan 326
Bhuiyan, Alauddin 117
Bootz, Friedrich 51
Bousse, Alexandre 167
Bressmann, Tim 151
Cazoulat, Guillaume 142
Chen, Songcan 241
Cheng, Bo 241
Chiusano, Gabriele 67
Cho, Zang-Hee 100
Chong, Vincent 134
Chung, Moo K. 225
Chupin, Marie 201
Clemens, Mark G. 209
Colliot, Olivier 201
Comaniciu, Dorin 91, 282
Cooper, Cyrus 184
Correa, Juan Carlos 142
Cuingnet, Rémi 201
Dai, Dai 193
Dam, Erik 109
Davatzikos, Christos 26
de Crevoisier, Renaud 142
Deroose, Christophe M. 233
Dikici, Engin 43
Ding, Hu 308
Drán, Gaël 142
Duncan, James S. 83
Eichhorn, Klaus W.G. 51
Feulner, Johannes 91
Filipovych, Roman 26
Fletcher, P. Thomas 59
Funka-Lea, Gareth 282
Gao, Liang 317
George, Jose 233
Glaunès, Joan Alexis 201
Gray, Katherine R. 159
Greiser, Andreas 282
Haigron, Pascal 142
Hamarneh, Ghassan 151
Hammers, Alexander 159
Hammon, Matthias 91
Hammoudi, Ahmad A. 317
He, Huiguang 193
Heckemann, Rolf A. 159
Hontani, Hidekata 257
Hornegger, Joachim 91, 282
Hou, Zengguang 193
Huang, Weimin 134
Huber, Markus B. 352
Hutton, Brian F. 167
Ionasec, Razvan Ioan 282
Javaid, M. Kassim 184
Jia, Hongjun 175
Jian, Bing 126
Keustersmans, Johannes 249
Khakabi, Sina 291
Kim, Minjeong 100, 175
Krol, Andrzej 352
Kumar, Ashok J. 265
Last, Carsten 51
Lee, Su-Lin 300
Lee, Tim K. 291
Leinsinger, Gerda 352
Leitner, François 18
Li, Fuhai 317
Li, Jiang 265
Li, Wei 100
Li, Yaohang 265
Liang, Jianming 273

- Liao, Shu 1
 Lillholm, Martin 10
 Liu, Wei 59
 Loeckx, Dirk 233
 Lu, Chao 83
 Lu, Wenmiao 335
 Lu, Xiaoguang 282
 Marques, Joselene 109
 Masson-Sibut, Agnès 18
 Massoud, Yehia 317
 McKenzie, Frederic D. 265
 McManigle, John 75
 Merrifield, Robert 300
 Morsillo, Nicholas 326
 Mueller, Edgar 282
 Mysling, Peter 10
 Nagarajan, Mahesh B. 352
 Nakib, Amir 18
 Nguyen, Nhat H. 209
 Nguyen, Uyen T.V. 117
 Nielsen, Mads 10
 Noble, J. Alison 35, 75, 184
 Norris, Eric 209
 Nyuys, Johan 233
 Orderud, Fredrik 43
 Ospina, Juan David 142
 Ourselin, Sébastien 167
 Pal, Christopher 326
 Papageorgiou, Aris 35
 Park, Laurence A.F. 117
 Pedemonte, Stefano 167
 Petersen, Kersten 10
 Petit, Eric 18
 Rahmatullah, Bahbibi 35
 Ramamohanarao, Kotagiri 117
 Resnick, Susan M. 26
 Rueckert, Daniel 159
 Sawada, Yoshihide 257
 Schlossbauer, Thomas 352
 Schmidt, Michaela 282
 Seo, Seongho 225
 Shen, Dinggang 1, 100, 175, 241, 344
 Shi, Yonghong 1
 Shin, Min C. 209
 Simon, Antoine 142
 Smeets, Dirk 249
 Son, Young-Don 100
 Song, Qi 308
 Sonka, Milan 308
 Staglianò, Alessandra 67
 Stojkovic, Branislav 308
 Suetens, Paul 233, 249
 Sun, Xiaoyan 265
 Suzuki, Kenji 360
 Tajbakhsh, Nima 273
 Tang, Lisa 151
 Tejpar, Sabine 233
 Thrall, Michael J. 317
 Tian, Qi 134
 Tran, Loc 265
 Tsymbal, Alexey 282
 Vandermeulen, Dirk 249
 Verhoek, Michael 75
 Verri, Alessandro 67
 Vinning, David 265
 Vitanovski, Dime 282
 Vogelstein, Joshua 193
 Vorperian, Houri K. 225
 Vunckx, Kathleen 233
 Wahl, Friedrich M. 51
 Wang, Ching-Wei 217
 Wang, Jihong 265
 Wang, Li 100
 Wang, Lichao 300
 Wang, Qian 175
 Wang, Yaping 175
 Wang, Zhimin 134
 Wang, Zhiyong 317
 Winkelbach, Simon 51
 Wismüller, Axel 352
 Wong, Stephen T.C. 317
 Wu, Guorong 100, 175
 Wu, Hong 273
 Wu, Tao 126
 Wu, Xiaodong 308
 Xiong, Wei 134
 Xu, Jian-Wu 360
 Xu, Jinhui 308
 Xu, Lei 308
 Xue, Wenzhe 273

- Yang, Guang-Zhong 300
Yaqub, Mohammad 75, 184
Yu, Cheng-Ping 217
Yurgelun-Todd, Deborah 59
- Zhang, Daoqiang 241, 344
Zhang, Kunlei 335
Zhou, Jiayin 134
Zhou, S. Kevin 91
Zhou, Xiang Sean 126

