Kanwal K. Bhatia
Herve Lombaert (Eds.)

LNCS 9487

Machine Learning Meets Medical Imaging

First International Workshop, MLMMI 2015
Held in Conjunction with ICML 2015
Lille, France, July 11, 2015, Revised Selected Papers

Lecture Notes in Computer Science 9487

Commenced Publication in 1973
Founding and Former Series Editors:
Gerhard Goos, Juris Hartmanis, and Jan van Leeuwen

Editorial Board

More information about this series at http://www.springer.com/series/7412

Kanwal K. Bhatia · Herve Lombaert (Eds.)

Machine Learning Meets Medical Imaging

First International Workshop, MLMMI 2015
Held in Conjunction with ICML 2015
Lille, France, July 11, 2015
Revised Selected Papers

Springer

Editors
Kanwal K. Bhatia
Imperial College
London
UK

Herve Lombaert
INRIA Sophia-Antipolis
Valbonne
France

ISSN 0302-9743 ISSN 1611-3349 (electronic)
Lecture Notes in Computer Science
ISBN 978-3-319-27928-2 ISBN 978-3-319-27929-9 (eBook)
DOI 10.1007/978-3-319-27929-9

Library of Congress Control Number: 2015957797

LNCS Sublibrary: SL6 – Image Processing, Computer Vision, Pattern Recognition, and Graphics

Printed on acid-free paper

This Springer imprint is published by SpringerNature
The registered company is Springer International Publishing AG Switzerland

Preface

This volume contains the proceedings of the First International Workshop on Medical Learning Meets Medical Imaging (MLMMI 2015) held on July 11, 2015, in Lille, France, in conjunction with the 32nd International Conference on Machine Learning (ICML 2015).

This workshop presented original methods and applications on machine learning in medical imaging. Developments in machine learning have opened up a wealth of novel opportunities in knowledge discovery, analysis, visualization, and reconstruction of medical image datasets. However, medical images also pose several particular challenges for standard approaches, for instance, lack of data availability (due to ethics or rarity of pathology), poor image quality (due to imaging or medical condition), or dedicated training requirements.

The workshop offered a unique opportunity to present and discuss the latest work on machine learning in medical imaging in the presence of both the machine learning and medical imaging communities. Innovative contributions addressed questions such as how to better exploit smaller datasets and understand the fundamentals of image spaces or generative models in order to improve training in machine learning methods. The workshop focused on theoretical aspects as well as on effective applications built on machine learning and all aspects of medical imaging.

The objective of this workshop was to bring together these specialties, to foster links, and to further communicate the specific needs and nuances of medical imaging to the machine learning community while exposing the medical imaging community to current trends in machine learning. We are also extremely grateful to the contributors of the first MLMMI workshop. We thank all authors who shared their latest findings, as well as the Program Committee members, and reviewers, who all achieved quality work in a very short time.

We also thank our keynote speakers, who kindly accepted our invitations: Bertrand Thirion, Research Director at Inria, France; John Ashburner, Professor in the Functional Imaging Laboratory at University College London; Marleen de Bruijne, Professor in the Erasmus Medical Center, in Rotterdam and in Computer Science at the University of Copenhagen; and Ben Glocker, Lecturer at Imperial College London. Special thanks go to the generous sponsors of our workshop: the Microsoft Research – Inria Joint Centre, in Palaiseau, France, and Imperial College London, UK. Congratulations go to Jonathan Young, who received the best paper award, kindly sponsored by NVidia Inc.

July 2015

Kanwal K. Bhatia
Herve Lombaert

Organization

Organizing Committee

Kanwal K. Bhatia	Imperial College London, UK
Herve Lombaert	Microsoft Research – Inria Joint Centre, France

Program Committee

Eugene Chang	University of Sheffield, UK
Laurent Charlin	Columbia University, USA
Pankaj Daga	University College London, UK
Claire Donoghue	University of Manchester, UK
Samuel Kadoury	Ecole Polytechnique de Montreal, Canada
Ender Konokoglu	MGH/Harvard, USA
Loic Le Folgoc	Inria, France
Karim Lekadir	Universitat Pompeu Fabra, Spain
Marco Lorenzi	University College London, UK
Jan Margeta	Inria, France
Ozan Oktay	Imperial College London, UK
Anil Rao	University College London, UK
David Rivest-Henault	CSIRO Brisbane, Australia
Matthias Schneider	ETH Zurich, Switzerland
Tong Tong	Imperial College London, UK
Raphael Sznitman	University of Bern, Switzerland
Maria Zuluaga	University College London, UK

Additional Reviewers

Salim Arslan
Eugene Chang
Laurent Charlin
Pankaj Daga
Claire Donoghue
Vikash Gupta
Samuel Kadoury
Ender Konokoglu
Loic Le Folgoc
Karim Lekadir
Marco Lorenzi
Jan Margeta
Ozan Oktay
Anil Rao
David Rivest-Henault
Matthias Schneider
Raphael Sznitman
Tong Tong
Maria Zuluaga

Sponsoring Institutions

Microsoft Research – Inria Joint Centre, Palaiseau, France
Imperial College London, London, UK
NVidia Inc., Santa Clara, CA

Contents

Motion

Retrospective Motion Correction of Magnitude-Input MR Images 3
Alexander Loktyushin, Christian Schuler, Klaus Scheffler, and Bernhard Schölkopf

Automatic Brain Localization in Fetal MRI Using Superpixel Graphs 13
Amir Alansary, Matthew Lee, Kevin Keraudren, Bernhard Kainz, Christina Malamateniou, Mary Rutherford, Joseph V. Hajnal, Ben Glocker, and Daniel Rueckert

Brain

Learning Deep Temporal Representations for fMRI Brain Decoding 25
Orhan Firat, Emre Aksan, Ilke Oztekin, and Fatos T. Yarman Vural

Modelling Non-stationary and Non-separable Spatio-Temporal Changes in Neurodegeneration via Gaussian Process Convolution 35
Marco Lorenzi, Gabriel Ziegler, Daniel C. Alexander, and Sebastien Ourselin

Improving MRI Brain Image Classification with Anatomical Regional Kernels 45
Jonathan Young, Alex Mendelson, M. Jorge Cardoso, Marc Modat, John Ashburner, and Sebastien Ourselin

Computer Aided Diagnosis

A Graph Based Classification Method for Multiple Sclerosis Clinical Forms Using Support Vector Machine 57
Claudio Stamile, Gabriel Kocevar, Salem Hannoun, Françoise Durand-Dubief, and Dominique Sappey-Marinier

Classification of Alzheimer's Disease Using Discriminant Manifolds of Hippocampus Shapes 65
Mahsa Shakeri, Hervé Lombaert, and Samuel Kadoury

Transfer Learning for Prostate Cancer Mapping Based on Multicentric MR Imaging Databases 74
Rahaf Aljundi, Jérôme Lehaire, Fabrice Prost-Boucle, Olivier Rouvière, and Carole Lartizien

Segmentation

Feature-Space Transformation Improves Supervised Segmentation Across Scanners 85
Annegreet van Opbroek, Hakim C. Achterberg, and Marleen de Bruijne

Discriminative Dimensionality Reduction for Patch-Based Label Fusion. 94
Gerard Sanroma, Oualid M. Benkarim, Gemma Piella, Guorong Wu, Xiaofeng Zhu, Dinggang Shen, and Miguel Ángel González Ballester

Author Index 105

Motion

Retrospective Motion Correction of Magnitude-Input MR Images

Alexander Loktyushin(✉), Christian Schuler, Klaus Scheffler, and Bernhard Schölkopf

Max Planck Institute for Intelligent Systems,
Spemannstraße 41, 72076 Tübingen, Germany
aloktyus@tuebingen.mpg.de

Abstract. There has been a considerable progress recently in understanding and developing solutions to the problem of image quality deterioration due to patients' motion in MR scanners. Retrospective methods can be applied to previously acquired motion corrupted data, however, such methods require complex-valued raw volumes as input. It is common practice, though, to preserve only spatial magnitudes of the medical scans, which makes the existing post-processing-based approaches inapplicable. In this work, we make first humble steps towards solving the problem of motion-related artifacts in magnitude-only scans. We propose a learning-based approach, which involves using large-scale convolutional neural networks to learn the transformation from motion-corrupted magnitude observations to the sharp images.

1 Introduction

The problem of image quality degradation due to subjects' motion during the acquisition remains one of the most important problems in MRI with no universally accepted solution [1]. Motion causes ghosting and blurring of the images, and even a millimeter displacements of the head position during the scan can make the image non-diagnostic. High resolution scans are particularly prone to motion because they require long acquisition times, and since high-frequency details of the image can be distorted even by sub-millimeter movements.

Over the last decades a multitude of approaches have been proposed. They can be broadly categorized into two classes: prospective and retrospective approaches. Prospective methods [2] aim at solving the problem of motion online during the acquisition by constantly adjusting the spatial encoding gradients to accommodate for the position change of the scanned object. This requires real-time motion tracking, which can be achieved with the use of tracking cameras [3], active markers [4] or the scanner itself [5]. Although there has been a lot of progress recently in improving prospective methods, they are still mostly used in research facilities and not in clinical practice.

Retrospective approaches, on the other hand, attempt to correct for motion artifacts once the data is acquired, which means that they can be potentially applied to any scan that was acquired in the past. An important subclass of

K.K. Bhatia and H. Lombaert (Eds.): MLMMI 2015, LNCS 9487, pp. 3–12, 2015.
DOI: 10.1007/978-3-319-27929-9_1

such methods are autofocusing-based approaches, which are purely data-driven. The origin of the autofocusing methods [6] can be traced to the early attempts in solving image denoising and deblurring problems. The idea of autofocusing is to search for a set of motion parameters, which is used to invert the effects of motion in the image. The motion trajectory associated with the lowest value of the image quality metric evaluated in the spatial domain is then selected. Typically, the entropy of the image gradients is used as an image metric. Such approaches rely on the fact that the effects of motion are perfectly invertible in case of translational motion, and approximately invertible in case of rotations by small angles (typically below 10 degrees). In case of rotations, k-space regridding is performed in order to correct for motion.

An important aspect of retrospective methods is that contrasted with prospective approaches they can be also applied to correct for non-rigid motion [7]. However, retrospective approaches suffer from long computation times when it comes to correction, which hinders their use in the time-critical environment of medical facilities. This shortcoming was addressed in recent work, where fast correction of both rigid [8] and non-rigid [9] motion was shown to be possible (seconds scale for rigid motion, and minutes for non-rigid motion). However, to our knowledge, all current retrospective methods require complex-valued raw data as input to do the correction, while most medical images are usually stored in DICOM format with only the spatial magnitude being preserved. The problem with the magnitude images is that the effects of motion can not be inverted in one step closed form solution.

It has recently been shown that it is possible to use large scale deep neural networks to achieve state of the art results in denoising [10], non-blind [11] and blind deblurring problems [12]. In this proof-of-concept paper we make early steps into the exploration of the problem of motion correction of magnitude-only medical scans, which we show to be similar to the image deblurring problem. We use large scale convolutional neural networks [13] to learn the mapping from motion corrupted magnitude-only scans to artifact-free images. We benefit from the fact that the exact model of the motion transformation is known, which makes it possible to generate an arbitrary amount of motion-corrupted data for training thus avoiding the problem of overfitting.

2 Methods

We first describe the model of the image degradation due to motion, and then formalize and discuss the problem of correction of the magnitude-only images. We then propose an iterative solution to the special case of the problem, namely that the latent image is assumed to be real-valued and the motion transformation to be known. We then take the next step and consider the more realistic problem of unknown motion degradation, and complex-valued latent images, which we address with a learning-based approach. Please note that in this preliminary work we deal with a simple version of the problem, where the images are assumed to be 2D, the motion to have no rotational components, and the data to be acquired with a single channel coil.

2.1 Models of the Image Degradation Due to Motion

There are two fundamental ways the effects of the motion of the scanned object on the image can be described. Lets first introduce the necessary formalism. Let $\mathbf{u} \in \mathbb{C}^n$ be an unknown sharp 2D image of size $n = n_1 \cdot n_2$ pixels, $\tilde{\mathbf{A}}_{\boldsymbol{\theta}_t}$ a rigid motion transformation matrix, $\boldsymbol{\theta}_t \in \mathbb{R}^2 \times [0, 2\pi)$ a vector with translation and rotation motion parameters at time t, T the length of the acquisition (total number of time points), and $\mathbf{F} \in \mathbb{C}^{n \times n}$ the orthonormal Fourier matrix. In this preliminary work, we assume that there is no noise. There are two ways the effects of motion can be modeled:

Spatial domain model:

The image acquisition with motion can be written as a linear process

$$\mathbf{y} = \sum_{t=1}^{T} \mathrm{diag}(\mathbf{m}_t)\mathbf{F}\tilde{\mathbf{A}}_{\boldsymbol{\theta}_t}\mathbf{u} \in \mathbb{C}^n,\ \mathbf{1} = \sum_{t=1}^{T} \mathbf{m}_t,$$

where $\mathbf{m}_t \in [0, 1]^n$ is a diagonal masking matrix selecting the segment in k-space acquired by the scanner at time t.

Fourier domain model:

Let $\mathbf{A}_{\boldsymbol{\theta}}$ be an operator, such that $\mathbf{F}\tilde{\mathbf{A}}_{\boldsymbol{\theta}} = \mathbf{A}_{\boldsymbol{\theta}}\mathbf{F}$. Then acquisition in k-space can then be written as

$$\mathbf{y} = \sum_{t=1}^{T} \mathrm{diag}(\mathbf{m}_t)\mathbf{A}_{\boldsymbol{\theta}_t}\mathbf{F}\mathbf{u} \in \mathbb{C}^n.$$

This representation allows for a more compact description of the problem, where we can now use a single matrix $\mathbf{A}_{\boldsymbol{\Theta}}$ to describe the cumulative effects of motion on the image

$$\mathbf{y} = \mathbf{A}_{\boldsymbol{\Theta}}\mathbf{F}\mathbf{u} \in \mathbb{C}^n,\ \mathbf{A}_{\boldsymbol{\Theta}} := \begin{bmatrix} [\mathbf{A}_{\boldsymbol{\theta}_1}]_{\mathbf{m}_1} \\ [\mathbf{A}_{\boldsymbol{\theta}_2}]_{\mathbf{m}_2} \\ \vdots \\ [\mathbf{A}_{\boldsymbol{\theta}_T}]_{\mathbf{m}_T} \end{bmatrix} \in \mathbb{C}^{n \times n},$$

where the vector $\boldsymbol{\Theta} = [\boldsymbol{\theta}_1, .., \boldsymbol{\theta}_T] \in \mathbb{R}^{2T} \times [0, 2\pi)^T$ now contains the motion parameters from the entire trajectory (all repetitions).

In this work, we restrict ourselves to the case where the scanned object performs purely translational motions. In this case the matrix $\mathbf{A}_{\boldsymbol{\Theta}}$ is diagonal, and its entries relate to the frequency-dependent phase ramps with slopes equal to spatial displacements. An important insight is that in this case the degradation due to motion can be written as a convolution in spatial domain $\mathbf{y} = \mathbf{F}(\mathbf{a}_{\boldsymbol{\Theta}} \star \mathbf{u})$, with the kernel $\mathbf{a}_{\boldsymbol{\Theta}}$. Such a convolution kernel has very special properties: it has a large support in spatial domain (basically the size of the image), it is complex-valued, and its power spectrum is unit-valued in all frequencies.

2.2 Magnitude-Only Image Problem

In many practical cases the raw observation $\mathbf{y}$ is not available, and only the magnitude of the image in spatial domain $\mathbf{z} = |\mathbf{a}_\Theta \star \mathbf{u}|$ is what is left from the scan. As a trivial case, yet containing the crucial aspects of the problem, we can consider a two-element image $\mathbf{u} \in \mathbb{C}^2$ and a two-element kernel $\mathbf{a} \in \mathbb{C}^2$, with entries equal to $\mathbf{u} = [\hat{u}_1 e^{i\beta_1}, \hat{u}_2 e^{i\beta_2}]$ and $\mathbf{a} = [\hat{a}_1 e^{i\alpha_1}, \hat{a}_2 e^{i\alpha_2}]$ respectively. Computing the convolution and taking the complex modulus we obtain for the first element z_1:

$$\begin{aligned} z_1 &= (\hat{u}_1\hat{a}_1 e^{i(\alpha_1+\beta_1)} + \hat{u}_2\hat{a}_2 e^{i(\alpha_2+\beta_2)})(\hat{u}_1\hat{a}_1 e^{-i(\alpha_1+\beta_1)} + \\ &+ \hat{u}_2\hat{a}_2 e^{-i(\alpha_2+\beta_2)}) \\ &= (\hat{u}_1\hat{a}_1 + \hat{u}_2\hat{a}_2)^2 + 2\hat{u}_1\hat{a}_1\hat{u}_2\hat{a}_2(\cos(\alpha_1 + \beta_1 - \\ &- \alpha_2 - \beta_2) - 1). \end{aligned}$$

We see that the degraded result is not a convolution anymore, and even worse, it depends on the phase difference, which is lost. In case the phase variation is smooth over the image we can drop the term with phases, but this is not likely to be the case in real images.

2.3 Iterative Solution

Under the assumption that the image $\mathbf{u}$ is real-valued, and the convolution kernel $\mathbf{a}_\Theta$ is known, it is possible to use the following iterative procedure (see Algorithm 1), which is our first contribution to solving the problem.

Algorithm 1. Non-blind iterative magnitude image motion correction

Input: Corrupted magnitude image $\mathbf{z}$; translational convolution kernel $\mathbf{a}_\Theta$
Output: Motion-corrected volume $\mathbf{u}$ in spatial domain
Initialize $\mathbf{u} \leftarrow \mathbf{z}$.
For $s \leftarrow 1, ..., N$ **do**

$$\mathbf{x} = \mathbf{z} \odot \exp(i \cdot \angle(\mathbf{a}_\Theta \star |\mathbf{u}|)).$$

$$\mathbf{u} \leftarrow \mathbf{F}^{\mathsf{H}} \frac{\mathbf{F}\mathbf{x}}{\mathbf{F}\mathbf{a}_\Theta}.$$

End

The idea behind this algorithm is to do a multiple reconvolution/deconvolution phase updates, while keeping the magnitude of the observation fixed. In our experiments, we observed that 100 iterations were sufficient. Although the proposed iterative procedure has certain interesting theoretical aspects it is of little use when it comes to motion correction of medical data, where the convolution kernel $\mathbf{a}_\Theta$ is not known, and the latent image is complex-valued.

2.4 Learning-Based Approach

Inspired by the recent success in using large scale learning-based approaches to solving the standard computational photography problems such as denoising and blind deconvolution, we decided to try using neural networks to learn the transformation that undoes the effects of motion in magnitude images. The network is given the motion-corrupted magnitude observation $\mathbf{z}$ as input, and the target is the magnitude of the latent image $\mathbf{u}$. We choose to predict just the magnitude of the image (containing all the necessary medical information), since it simplifies the inference problem. To generate the data we used a 3D scan of the brain from a healthy volunteer (acquired with MPRAGE sequence), and split it into 70 two-dimensional slices. We then used 50 randomly-selected slices for training, and the remaining 20 slices for validation.

On each training iteration we generated random translational motion trajectories $\boldsymbol{\Theta}$ with maximum displacements of up to five pixels. We then used the forward model of translational motion parameterized by randomly generated trajectories to simulate the effects of motion in slice images. On each training iteration the random slice was picked and an input/target pair was generated. Thus on each training iteration the network always receives a new distinct input subject to the infinitely many possible variations in the motion trajectories.

Over the training process we also evaluated the training curve by generating the inputs based on validation slices and random motion. We computed the mean squared error between the predicted magnitudes and motion-free magnitudes. This way we could track the progress in the network training.

Although the input and the output layers of the network only deal with magnitudes, the underlying motion transformation that generates the magnitude observation from the complex-valued latent image is applied in complex domain. Since we can generate arbitrary many motion trajectories, we are unlikely to overfit the data.

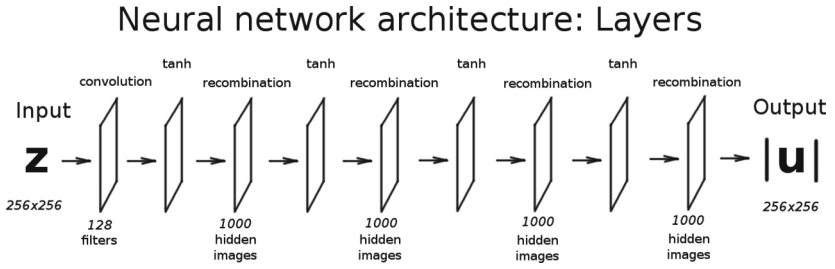


Fig. 1. The overview of the neural network architecture used for motion correction. The input to the network is the motion-corrupted magnitude observation $\mathbf{z}$, output is the magnitude of the motion-free original image $\mathbf{u}$.

The architecture of the network that we used is outlined in Fig. 1. The first layer is a convolutional layer with 128 large filters of size 192×128 pixels. We have observed that using large filters is necessary for good prediction results,

which has to do with the fact that the motion kernels $\mathbf{a}_{\boldsymbol{\Theta}}$ have large spatial support. We used our own custom-built C++/CUDA-based neural network implementation with circular convolution operations. The convolution layer outputs 128 hidden image planes, which are then passed through a non-linear tanh layer, and get recombined in the subsequent linear layer to 1000 new hidden images. The recombination interleaved with the tanh coordinate-wise operation is then repeated three times. We used $3 \cdot 10^5$ training iterations, the Adadelta parameter update rule [14] with an adaptive learning rate, and MSE criterion as a loss function.

3 Results

We start with an experiment on the simulated data with a simpler non-blind problem for which an iterative solution is sufficient. We then show the results of the second experiment, where we consider the harder blind problem and use the learning-based approach to solve it.

3.1 Iterative Reconstruction

In Fig. 2 we show the results that were obtained using the iterative procedure (Algorithm 1). In this case, the complex-valued convolution kernel is provided to the algorithm as input, and the latent image is assumed to be real-valued.

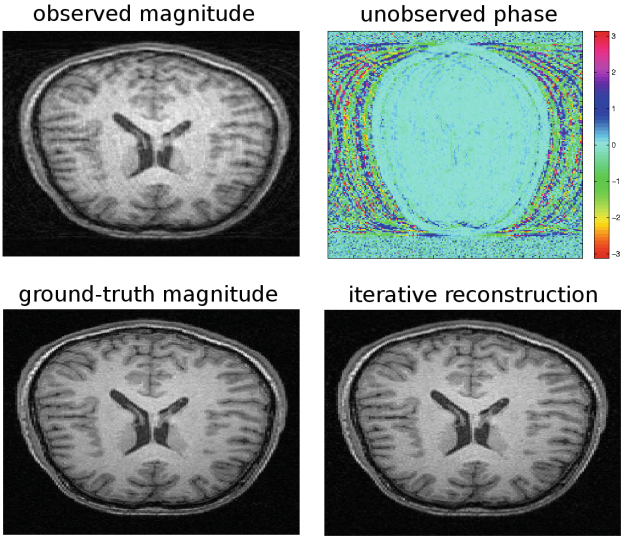


Fig. 2. The results of motion correction obtained with the iterative approach. The complex-valued kernel is assumed to be known, and the image to be real-valued.

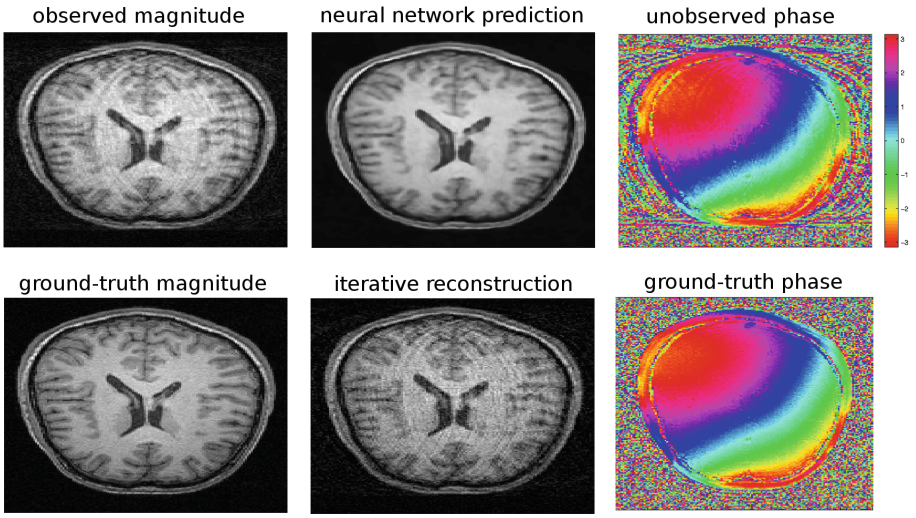


Fig. 3. Neural network-based prediction. Here the convolution kernel is not known and the latent image is assumed to be complex-valued. Also shown is the result obtained with the iterative procedure (having access to the kernel).

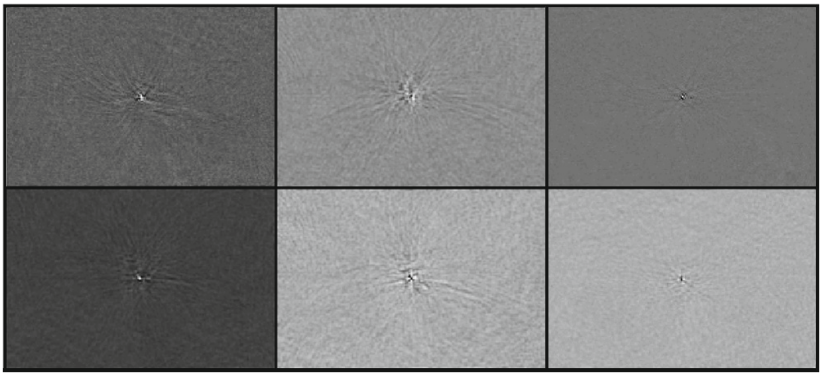


Fig. 4. Four filters (each of size 192×128) extracted from the convolutional layer of the network after training.

3.2 Learning-Based Solution

In a realistic scenario the motion trajectory and thus the kernel are not known. Furthermore, MR images have a non-trivial spatial phase, which depends on the sequence, magnetic properties of the scanned object and shims. In Fig. 3 we show motion correction results produced by the neural network. As input, the network received just the magnitude observation (from the validation slice dataset). Although some residual blurring can be seen, the network-based prediction was successful at removing the ghosting motion artifacts. In the same

figure we also plot the results obtained with the iterative procedure, for which the blurring kernel was provided as input. In this case, the iterative procedure fails to remove the artifacts because the latent image is complex-valued (we show the spatial phase in the figure).

In Fig. 4 we plot six filters (out of the 128 filters from the first convolution layer) that were learned by the network. The filters display an intricate structure with a large spatial extent, which justifies using large filters in convolution layer.

4 Discussion

Retrospective motion correction of magnitude-only MR scans is in many ways similar to the well-studied problem of blind deconvolution. However, it also poses some unique challenges due to the special structure of the motion degradation kernel. In this paper, we explore the capability of a large-scale convolutional neural network to confront these challenges. Compared to the neural network designs commonly used to solve image denoising/deblurring problems our network has unusually large convolution filters in the first layer. Experimenting with different sizes of the filters we observed that larger filters allow for more improvement in the visual quality of reconstructed images. A possible explanation for this observation is that the translational motion point spread function has a large spatial support, which makes the problem highly non-local and necessitates the use of large receptive fields in the first layer. Also, the neural networks used for deblurring/denoising rely on splitting the image into multiple small patches, and training on the patches, not the images. In our case, this was not to be feasible due to the above mentioned problem of non-local image degradation. We use large scale neural networks and not the other learning based approaches since convolutional neural networks have a structure (local receptive fields) well-suited for the medical image data. Furthermore, such networks can be efficiently implemented to run on modern GPUs, which allows for efficient computations and reasonable training times, while having millions of neurons/parameters.

Having a network of such a large capacity can make the training prone to overfitting. We address this problem by using the forward model of the motion degradation to generate an arbitrary amount of the training input/target data, which is affected by random translational motions. However, we use just 70 slices from a single 3D scan for training and validation. The next research goal would involve constructing a large database of complex-valued scans of different human subjects in order to have sufficiently many examples of the different brain morphologies. Still, it is yet far from clear how many medical scans are needed for training in order to be able to generalize to all practical variations of the medical data, where the intensity of the image is subject to the pulse sequence parameters and coil configuration. It remains to be shown if it is possible to have a good generalization over different image contrasts. The bottom line is that in order to make the presented approach practical sufficiently large and representative dataset is needed. The problem of the scarcity of medical data is well-known in the medical image processing community, and here we are furthermore hindered by the fact that we need raw datasets to simulate the motion.

In this work, we only do simulation-based experiments, where we correct the motion on artificially generated data. In real data there are both translational and rotational components of motion. Rotational motion is no longer a convolution (to some extent it can be approximated as a spatially-variant convolution), which makes the problem even more challenging, and it remains to be shown that using the neural networks is an adequate solution for this problem too. We further assumed that the data was acquired only from a single channel coil – an assumption that greatly simplifies the problem. Learning-based motion correction of sum-of-squares combined multi-coil data is another topic for future research.

We admit that the motion model we assume (purely translational motion) will be insufficient for many practical cases of motion, where large rotations (around the point where the head touches the pillow) can happen. The main purpose of this work was rather to probe the possibility of using learning-based approaches for solving magnitude-only image correction problem in a simplified setting. Still, we expect the translation-only correction method to be able to improve the image quality in the images that are affected by weak motion.

Since we have the ground-truth available, it is possible to use the objective numerical estimates of the image quality improvement. This is useful when comparing reconstruction results against the alternative approaches. However, to our knowledge no other method capable of correction of magnitude-only data exists. We thus prefer not to provide the numerical estimates of the image quality improvement, because such estimates can be misleading, especially if the plain MSE criterion is used.

To summarize, in this work we have attempted to cope with the problem of motion correction of magnitude-only images with the use of large scale neural networks. Our preliminary results indicate the potential of the proposed approach for solving this problem.

References

1. Zaitsev, M., Maclaren, J., Herbst, M.: Motion artifacts in MRI: a complex problem with many partial solutions. Journal of Magnetic Resonance Imaging (2015)
2. Maclaren, J., Herbst, M., Speck, O., Zaitsev, M.: Prospective motion correction in brain imaging: a review. Magn. Reson. Med. **69**(3), 621–636 (2012)
3. Zaitsev, M., Dold, C., Sakas, G., Hennig, J., Speck, O.: Magnetic resonance imaging of freely moving objects: prospective real-time motion correction using an external optical motion tracking system. Neuroimage **31**, 1038–1050 (2006)
4. Ooi, M.B., Krueger, S., Thomas, W.J., Swaminathan, S.V., Brown, T.R.: Prospective real-time correction for arbitrary head motion using active markers. Magn. Reson. Med. **62**, 943–954 (2009)
5. van der Kouwe, A.J.W., Benner, T., Dale, A.M.: Real-time rigid body motion correction and shimming using cloverleaf navigators. Magn. Reson. Med. **56**, 1019–1032 (2006)
6. Atkinson, D., Hill, D., Stoyle, P., Summers, P., Keevil, S.: Automatic correction of motion artifacts in magnetic resonance images using an entropy focus criterion. IEEE Trans. Med. Imaging **16**(6), 903–910 (1997)

7. Cheng, J.Y., Alley, M.T., Cunningham, C.H., Vasanawala, S.S., Pauly, J.M., Lustig, M.: Nonrigid motion correction in 3D using autofocusing with localized linear translations. Magn. Reson. Med. **68**(6), 1785–1997 (2012)
8. Loktyushin, A., Nickisch, H., Pohmann, R., Schölkopf, B.: Blind retrospective motion correction of MR images. Magnetic Resonance in Medicine (2013). doi:10.1002/mrm.24615. (Epub ahead of print)
9. Loktyushin, A., Nickisch, H., Pohmann, R., Schölkopf, B.: Blind multirigid retrospective motion correction of MR images. Magn. Reson. Med. **73**(4), 1457–1468 (2015)
10. Burger, H.C., Schuler, C.J., Harmeling, S.: Image denoising with multi-layer perceptrons, part 1: comparison with existing algorithms and with bounds. CoRR abs/1211.1544 (2012)
11. Schuler, C.J., Burger, H.C., Harmeling, S., Schölkopf, B.: A machine learning approach for non-blind image deconvolution. In: Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2013, pp. 1067–1074. IEEE Computer Society, Washington, DC (2013)
12. Schuler, C.J., Hirsch, M., Harmeling, S., Schölkopf, B.: Learning to deblur. CoRR abs/1406.7444 (2014)
13. Le Cun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
14. Zeiler, M.D.: ADADELTA: an adaptive learning rate method. CoRR abs/1212.5701 (2012)

Automatic Brain Localization in Fetal MRI Using Superpixel Graphs

Amir Alansary[1(✉)], Matthew Lee[1], Kevin Keraudren[1], Bernhard Kainz[1,2], Christina Malamateniou[2], Mary Rutherford[2], Joseph V. Hajnal[2], Ben Glocker[1], and Daniel Rueckert[1]

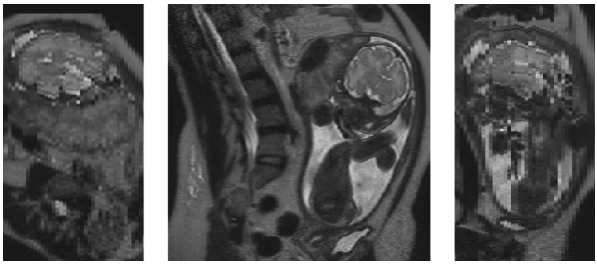
[1] Department of Computing, Imperial College London, London, UK
{a.alansary14,matthew.lee13,kevin.keraudren10,
b.kainz,b.glocker,d.rueckert}@imperial.ac.uk
[2] Division of Imaging Sciences, King's College London, London, UK
{christina.malamateniou,mary.rutherford,jo.hajnal}@kcl.ac.uk

Abstract. Fetal MRI is emerging as an effective, non-invasive tool in prenatal diagnosis and pregnancy follow-up. However, there is a significant variability of the position and orientation of the fetus in the MR images. This makes these images more difficult to analyze and interpret compared to standard adult MR imaging, which standardized anatomical imaging aligned planes. We address this issue by automatic localization of the fetal anatomy, in particular, the brain which is a structure of interest for many fetal MRI studies. We first extract superpixels followed by the computation of a histogram of features for each superpixel using bag of words based on dense scale invariant feature transform (DSIFT) descriptors. We construct a graph of superpixels and train a random forest classifier to distinguish between brain and non-brain superpixels. The localization framework has been tested on 55 MR datasets at gestational ages between 20–38 weeks. The proposed method was evaluated using 5-fold cross validation achieving a 94.55 % brain detection accuracy rate.

1 Introduction

Fetal magnetic resonance imaging (MRI) has significantly improved in the last two decades, and is emerging as a novel, non-invasive tool for diagnosis and planing of surgical interventions. It provides higher contrast and larger field-of-view than ultrasound. Thus, it provides better structural information of the different fetal organs such as the brain, spine and body. Fetal brain localization is important for assessing the fetal brain development and maturation. It is also the primary step for most of the current automatic motion correction techniques for fetal MRI [10]. Recently, fetal brain detection has been used as a landmark to extract the other fetal organs [11]. Problems that hinder the design of automated image analysis tools for fetal MRI usually arise from: (a) the high variability in shape, size, orientation, and anatomical configuration of the fetus; (b) intensity non-uniformities (bias artifacts); (c) partial volume effects; and (d) motion artifacts caused by the unconstrained fetal motion (see Fig. 1).

K.K. Bhatia and H. Lombaert (Eds.): MLMMI 2015, LNCS 9487, pp. 13–22, 2015.
DOI: 10.1007/978-3-319-27929-9_2



(a) axial (b) coronal (c) sagittal

Fig. 1. Three orthogonal cutting planes through a stack of fetal MRI images. The quality of the in-plane (coronal) slices is not affected by motion, however, there are inter-slice artifacts appear in the out-of-plane views (axial and sagittal).

Related work: Fetal MRI is an emerging field of research, with little work focused on fully automatic processing of these datasets. In [2], 3D template matching is used to detect the eyes, enabling a subsequent 2D/3D graph-cut segmentation that extracts the brain. This approach is based on 3D templates and lacks the flexibility necessary to deal with motion artifacts as well as fetal abnormalities. The methods proposed in [9] and [12] address the variability of fetal MRI through machine learning. In [9], a Random Forest (RF) classifier first distinguishes between maternal and fetal tissues before classifying different tissues of the fetal head, while [12] combines prior knowledge of the fetal size with maximally stable extremal regions (MSER) detection and a bag-of-words model.

Contribution: In this paper we propose a fully-automated framework for localizing the fetal brain in fetal MRI scans. Rather then working on individual pixels we make use of superpixels for a faster and more efficient detection algorithm. Because of the nature of superpixels that most likely represents the rigid regions in the image, using superpixels neighbors instead of pixel neighbors can reduce the effect of motion artifacts. Therefore, we have developed a new superpixel graphical model based on both spatial and intensity distances in 3D. The proposed localization framework achieves 94.55 % accuracy for the brain detection and 98.18 % prediction accuracy of the center of the brain. The proposed approach does not require landmarks as in [9] or prior information such as the gestational age of the fetus as in [12].

2 Method

The proposed approach for the automatic localization of the fetal brain consists of four main steps as shown in Fig. 2. The input data of our system are 3D fetal MRI datasets. The first step is to decompose each 2D slice into $\{N_1, ..., N_s\}$ superpixels in order to minimize the local redundancy in the input data. By clustering and constructing a single descriptor for each superpixel we reduce the impact of noise on each descriptor whilst preserving homogeneous regions that

are likely belong to the same anatomical region. The second step is to calculate image descriptors for each pixel and then aggregate them into one histogram h_i for each i-superpixel. The third step is to build superpixel graphs based on each superpixel's neighbors. Then each superpixel's histogram is normalized with its neighbors in the graphical model. During the fourth step, we use a random forest to generate a probability map of the brain for every superpixel. Finally, this probability map is refined using another auto-context classifier followed by selecting the largest 3D component.

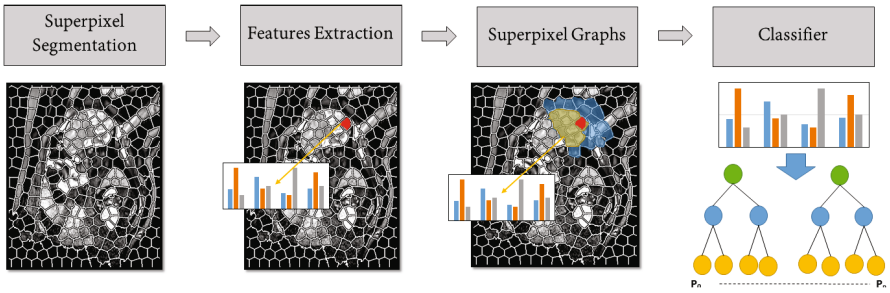
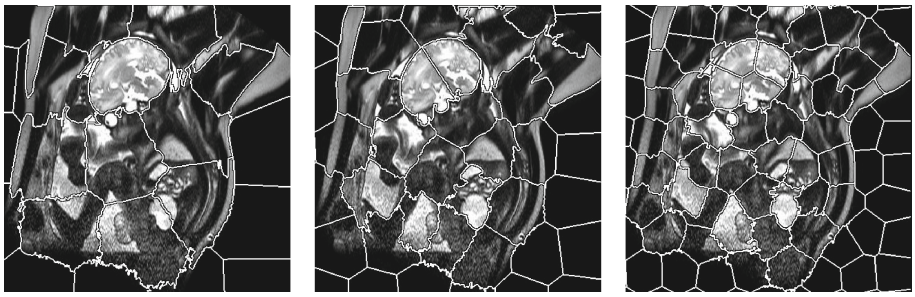


Fig. 2. The automatic localization of the fetal brain framework.

Superpixels: Superpixels are a popular unsupervised image segmentation technique that clusters image pixels into groups of pixels based on the correlation between each pixel and its neighbors. In the literature [1], many superpixel techniques have been shown to produce 'good' segmentation results. However, which properties of superpixels are important depends on the application. Medical image analysis can be computationally expensive when compared to normal image analysis due to the size of the data, namely because MRI scans are 3D volumes as opposed to 2D images.

Most of the current superpixel segmentation techniques have been proposed for 2D images. In our work, we have chosen the simple linear iterative clustering (SLIC) technique [1], which is fast to compute while achieving a good segmentation quality (as shown in Fig. 3) with lower computational cost so that the method scales well when processing the many slices of a volume. SLIC segments pixels into compact and nearly uniform superpixels. Superpixels are applied in 2D (not 3D) because of the fetal motion that results in 2D misaligned slices. Choosing the right number of superpixels for each 2D slice is challenging. Thus, we have modified the ad-hoc heuristics proposed by [8] to optimize superpixels for fetal MRI. We have weighted the rule with a constant factor k, where $k \in \mathbb{R}_{>0}$ and is chosen depending on n, the total number of pixels in the 2D input image. Thus, the total number of superpixels $s = |\{N_1, ..., N_s\}|$ in a 2D slice has been calculated:

$$s \approx k \cdot \sqrt{(n/2)}. \tag{1}$$



(a) $s \approx 25$ (b) $s \approx 50$ (c) $s \approx 100$

Fig. 3. A cropped image of a 2D fetal MRI scan segmented at (a) 25-, (b) 50-, and (c) 100-superpixels.

Image descriptors: We build a bag of features using dense scale invariant feature transform (DSIFT) [16] descriptors. This is done by first computing SIFT [13] descriptors for each pixel in every 2D image in the training set at a fixed scale and orientation. A k-means clustering is then performed on these descriptors and their centers are used to form a dictionary of k words. When collecting DSIFT descriptors from pixels we can then find their closest matching word from this dictionary, aggregating the frequency of words in each superpixel N_i into one histogram h_i with k-bins. The k-dimensional histogram acts as the feature descriptor for each superpixel. However, the descriptor is constructed in such a way that only contains local information about the superpixel itself. This leads to a loss of the large-scale image context. Also, due to the nature of the superpixels, their histograms of features tend to be sparse. Most of the DSIFT descriptors within a superpixel are likely to be mapped to the same word.

Superpixel graphs: To overcome the problem of sparse descriptors for superpixels, we construct superpixel graphs using the distance between the centroids of superpixels as edges [7]. These edges are weighted based on both spatial and geodesic distances. We first compute the centroid c_i for each N_i superpixel. Then we identify r neighbors based on the similarity score between two superpixels N_i and N_o with centroids c_i and c_o. The similarity between two superpixels is defined:

$$f_r(N_i, N_o) = 1 - \frac{d(c_i, c_o)}{D}, \tag{2}$$

Here $d(\cdot, \cdot)$ is the Euclidean distance, and D is the length of the diagonal of the 2D image. This normalizes the score to be between $[0, 1]$. The closer $f_r(c_i, c_o)$ is to 1, the smaller is the distance between these two centroids. We next assign weights w_j for the extracted r neighbors candidates based on the geodesic distances between their centroids. The geodesic distance is estimated using:

$$d_r(N_i, N_o) = \sum_{j=1}^{m} |I(p_j) - I(p_{j-1})|\,, \tag{3}$$

Here $I(p_j)$ is the intensity of the pixel p_j, and m equals to the number of pixels located on the straight line between c_i and c_o. Next, the weights w_j are calculated by normalizing d_r, so that $w_j = 1$ when d_r is the lowest and $w_j = 0$ when d_r is the highest. Finally, the histograms of features for the extracted r superpixels are aggregated based on the calculated weights, and normalized using:

$$\tilde{h}_i = \frac{\sum_{j=0}^{r} w_j h_j}{\left\| \sum_{j=0}^{r} w_j h_j \right\|_{\ell 1}} \tag{4}$$

Here $h_0 = h_i$ or the histogram of the current superpixel in consideration. Using graphs of superpixels enables the proposed localization method to overcome the motion artifacts between the 2D slices by extracting the superpixel neighbors in 3D, which would be more difficult using graphs of individual pixels. This is because of the nature of superpixels that most likely represent rigid regions, see Fig. 4. By selecting both spatial and intensity neighbors, we increase the features for each extracted histogram instead of using sparse features. Consequently, we extend the features used in the machine learning to include the image context instead of using only the local information. Figure 4 shows the proposed graphical model in both 2D and 3D.

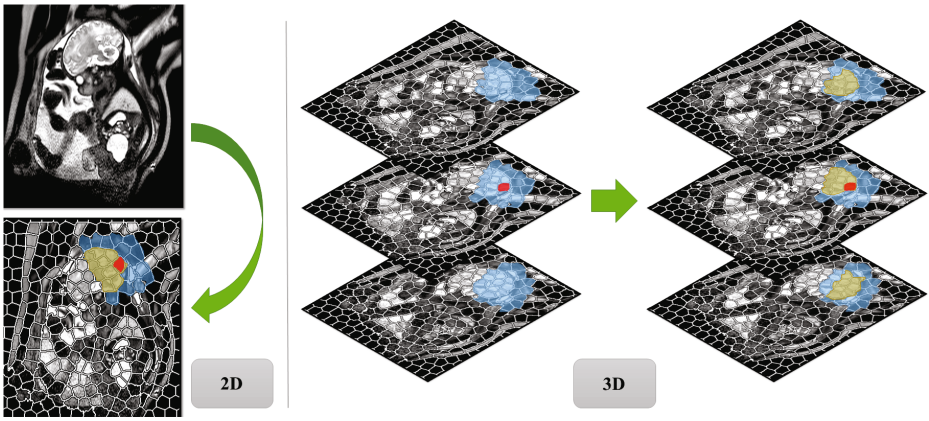


Fig. 4. The proposed superpixel graphical model in 2D and 3D. The red superpixel represents the current superpixel in consideration, the r-spatial neighbors are colored in blue, and the neighbors with higher weights (w_j) are colored in green(Color figure online).

Classification: We use the normalized feature vectors to train a two class random forest (RF) [4] to classify each superpixel as brain or non-brain. Our dataset has pixel-wise labels so we assign each superpixel N_i the class label l_i that corresponds to the label with the highest frequency inside the superpixel. After the initial classification we obtain a probability map and use this to train another second random forest along with 10 % of the most important features used for training the first classifier. This produces an auto-context classifier [15] that can increase the classification accuracy. The output volume is then filtered by finding the largest 3D component, which is the brain mask in this case, followed by a convex-hull extraction [3] to obtain a clean homogeneous segmentation.

3 Evaluation and Results

Data: The proposed framework has been tested on 55 fetal scans at gestational age between 20–38 weeks. Thirty subjects of these datasets are from normal fetuses and 25 datasets are from fetuses with intrauterine fetal growth restriction (IUGR). The data was acquired with a 1.5T Philips MRI system using single shot fast spin echo (ssFSE) sequences with voxel size $0.8398 \times 0.8398 \times 4\,\mathrm{mm}^3$. Ground truth labels were obtained by manual segmentation of the brain performed by expert observers.

Implementation: We perform mean and standard deviation normalization on the input scan intensities as a preprocessing stage for our proposed approach. We have adjusted the SLIC superpixel extraction used in [1] for generating superpixels that are optimized for fetal MRI data. The vlfeat library [16] was used for generating DSIFT features and the scikit-learn library [14] was used for the random forest classifier. The code was implemented using python and MatLab with the mex-c environment. We use $k = 5$ to determine the number of superpixels s. The number of neighbors selected for superpixel graphs r were set to 25 calculated for each $xy-$plane in three slices. In order to balance the positive and negative training samples for the classifier, we choose to restrict training to superpixels generated from a cropped volume around the brain by adding 25 % of the maximum brain diameter in the xy-plane. The prior knowledge that brains appear brighter in T2-scans allowed us to suppress some of the background pixels by thresholding any pixel less than 10 % of the maximum intensity value of the whole subject. All these parameters are chosen by experiment on a smaller test dataset.

Results: A 5-fold cross validation was used for evaluating our approach (11 test patients 44 training patients per fold). The random forest classifier achieved an average accuracy score of 96.17 % per a superpixel basis. We defined the detection accuracy, the extracted mask covering at least 70 % of the brain, similar to the definition presented in [9] but calculated for the whole 3D brain for simplification. The prediction accuracy of the centers of the segmented brains are measured by calculating the percentage of centers that lie inside the ground truth of the manually labeled brains. Brain coverage is measured by the percentage of the

Table 1. The accuracy of the proposed localization approach at different dictionary sizes $k = 50$, $k = 100$, $k = 400$ and $k = 800$.

	k = 50	k = 100	k = 400	k = 800
Brain detection (% subjects)	87.27	89.09	94.55	90.91
Center in brain (% subjects)	80	81.82	98.18	98.18
Brain coverage ($\mu \pm \sigma\%$)	85.59±30.51	85.9±28.2	90.03±16.63	90.54±15.17
Dice coefficient ($\mu \pm \sigma\%$)	61.07±26.94	63.22±25.47	71.96±19	73.62±15.9
Average train time (minutes)	31.55	50.25	256.09	424.93
Average test time (minutes)	4.63	4.85	6.31	8.36

manually-labeled brain that are covered by the segmented boundary box. We have also used the dice coefficient [6] in order to measure the segmentation accuracy of the proposed approach. Table 1 shows the accuracy of our localization approach at different dictionary sizes $k = 50$, $k = 100$, $k = 400$ and $k = 800$. These results shows that increasing the dictionary size or the histogram bins (sparsity) increases the dice accuracy. However, it also increases significantly the processing time of training and testing. These experiments were done using parallel processing on a CPU with 32-cores and 128 GB RAM (Fig. 5).

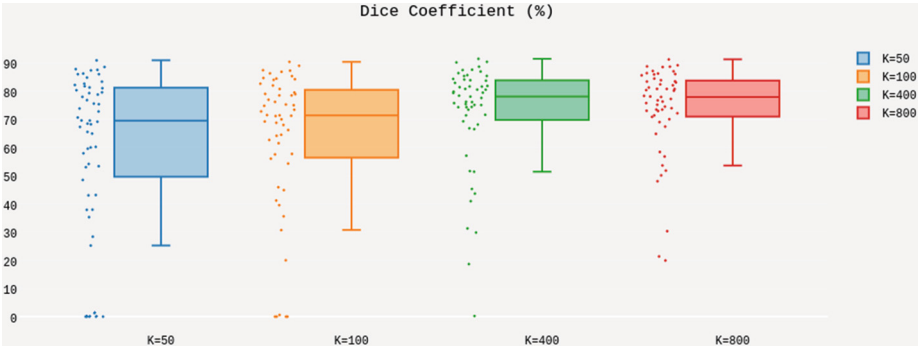
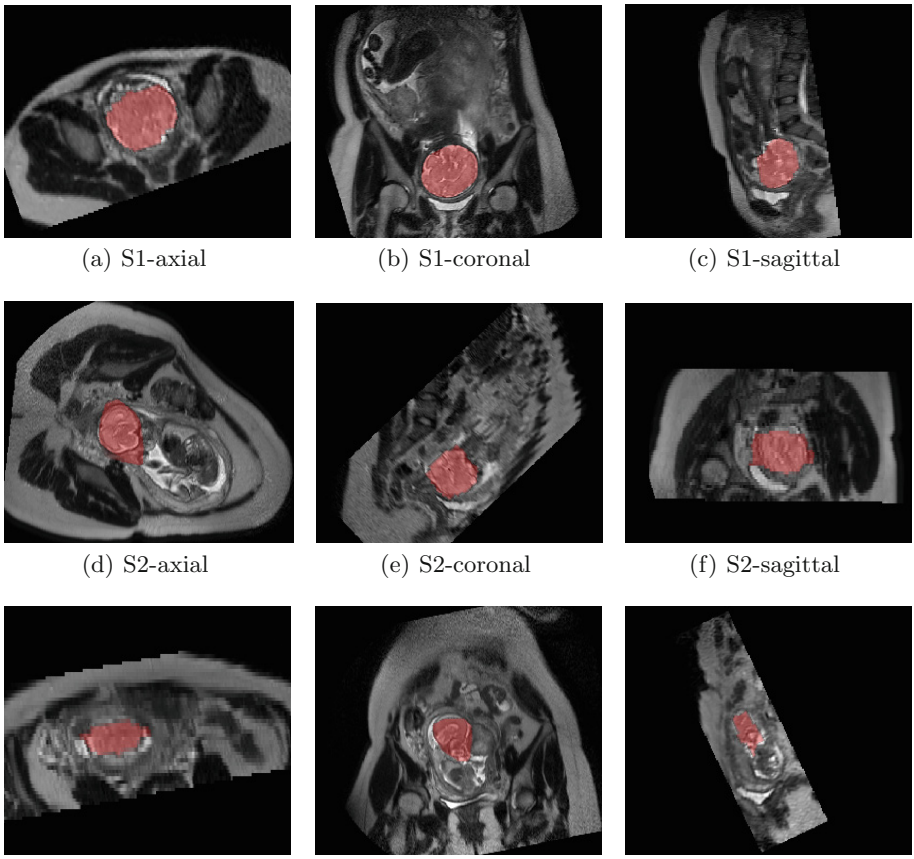


Fig. 5. The accuracy of the proposed localization approach at different dictionary sizes $k = 50$, $k = 100$, $k = 400$ and $k = 800$.

Our brain localization approach have achieved a 94.55 % detection accuracy. It also could detect the center of the brain with prediction accuracy 98.18 % of the test subjects while in [9] they achieved only 81 %, 78 % and 60 % using coronal, axial and sagittal training data. In addition, the proposed approach does not depend on the orientation of the acquired data and it does not use any previous landmarks as in [9]. [12] achieved 100 % detection accuracy of the brain; their method, however, requires previous information about the gestational age of the fetus to find the expected size of the brain. This information is later used to remove the outliers of the detected brain mask. Our proposed method has



(a) S1-axial (b) S1-coronal (c) S1-sagittal

(d) S2-axial (e) S2-coronal (f) S2-sagittal

(g) S3-axial (h) S3-coronal (i) S3-sagittal

Fig. 6. The segmented brain at different cross sections (axial, coronal, sagittal) for three different test subjects S1, S2 and S3. The dice coefficient of S1=91.22 %, S2=70.01 %, and S3=56.79 %.

an advantage to be generic and does not require any prior information. Figure 6 shows the segmented brain at different cross sections for three test subjects with different dice accuracy.

4 Discussion and Conclusion

We have developed an automatic framework for localizing the brain in fetal MRI scans using superpixel graphical models. Superpixels have enabled the proposed detection algorithm to be faster and more efficient than using pixels for classification. Also, extending the extracted features from the individual superpixels to include features from the neighbors using superpixel graphical models, have provided more information about the image context instead of using only the

local information. The evaluation results achieved 94.55 % accuracy for the brain detection, which shows the potential of extending the proposed approach using superpixel graphs to segment other fetal organs such as the heart, lung, and placenta. According to the recent studies [5], the placental functions affect the birth weight as the placenta controls the nutrients transmissions from the maternal to the fetal circulation. Moreover, the extracted brain can be used for developing automatic motion correction and registration techniques for fetal MRI.

Acknowledgments. Thanks for the volunteer subjects and radiographers from St. Thomas Hospital London for the image acquisitions. We used the Medical Imaging Interaction Toolkit (MITK) [17] to visualize some of the figures. Amir Alansary is supported by the Imperial College PhD Scholarship.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: SLIC superpixels compared to state-of-the-art superpixel methods. IEEE Trans. Pattern Anal. Mach. Intell. **34**(11), 2274–2282 (2012)
2. Anquez, J., Angelini, E.D., Bloch, I.: Automatic segmentation of head structures on fetal MRI. In: Proceedings of the IEEE International Symposium on Biomedical Imaging, pp. 109–112. IEEE Press (2009)
3. Barber, C.B., Dobkin, D.P., Huhdanpaa, H.: The quickhull algorithm for convex hulls. ACM Trans. Math. Softw. (TOMS) **22**(4), 469–483 (1996)
4. Breiman, L.: Random forests. Mach. Learn. **45**(1), 5–32 (2001)
5. Damodaram, M., Story, L., Eixarch, E., Patel, A., McGuinness, A., Allsop, J., Wyatt-Ashmead, J., Kumar, S., Rutherford, M.: Placental MRI in intrauterine fetal growth restriction. Placenta **31**(6), 491–498 (2010)
6. Dice, L.R.: Measures of the amount of ecologic association between species. Ecology **26**(3), 297–302 (1945)
7. Fulkerson, B., Vedaldi, A., Soatto, S.: Class segmentation and object localization with superpixel neighborhoods. In: 2009 IEEE 12th International Conference on Computer Vision, pp. 670–677. IEEE (2009)
8. Hair, J.F., Black, W.C., Babin, B.J., Anderson, R.E., Tatham, R.L.: Multivariate Data Analysis. Pearson, Prentice Hall, Upper Saddle River (2006)
9. Ison, M., Donner, R., Dittrich, E., Kasprian, G., Prayer, D., Langs, G.: Fully automated brain extraction and orientation in raw fetal MRI. In: Proceedings of the Workshop on Paediatric and Perinatal Imaging, MICCAI, vol. 12, pp. 17–24 (2012)
10. Keraudren, K., Kuklisova-Murgasova, M., Kyriakopoulou, V., Malamateniou, C., Rutherford, M., Kainz, B., Hajnal, J., Rueckert, D.: Automated fetal brain segmentation from 2D MRI slices for motion correction. Neuroimage **101**, 633–643 (2014)
11. Keraudren, K., Kainz, B., Oktay, O., Kyriakopoulou, V., Rutherford, M., Hajnal, J.V., Rueckert, D.: Automated localization of fetal organs in MRI using random forests with steerable features. In: Medical Image Computing and Computer-Assisted Interventation–MICCAI (2015) (in press)

12. Keraudren, K., Kyriakopoulou, V., Rutherford, M., Hajnal, J.V., Rueckert, D.: Localisation of the brain in fetal MRI using bundled SIFT features. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part I. LNCS, vol. 8149, pp. 582–589. Springer, Heidelberg (2013)
13. Lowe, D.G.: Object recognition from local scale-invariant features. In: The Proceedings of the Seventh IEEE International Conference on Computer Vision, vol. 2, pp. 1150–1157. IEEE (1999)
14. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., et al.: Scikit-learn: machine learning in python. J. Mach. Learn. Res. **12**, 2825–2830 (2011)
15. Tu, Z., Bai, X.: Auto-context and its application to high-level vision tasks and 3d brain image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **32**(10), 1744–1757 (2010)
16. Vedaldi, A., Fulkerson, B.: VLFeat: an open and portable library of computer vision algorithms. In: Proceedings of the International Conference on Multimedia, pp. 1469–1472. ACM (2010)
17. Wolf, I., Vetter, M., Wegner, I., Böttger, T., Nolden, M., Schöbinger, M., Hastenteufel, M., Kunert, T., Meinzer, H.P.: The medical imaging interaction toolkit. Med. Image Anal. **9**(6), 594–604 (2005)

Brain

Learning Deep Temporal Representations for fMRI Brain Decoding

Orhan Firat[1(✉)], Emre Aksan[1], Ilke Oztekin[2], and Fatos T. Yarman Vural[1]

[1] Middle East Technical University, Ankara, Turkey
orhan.firat@ceng.metu.edu.tr
[2] Koc University, Istanbul, Turkey

Abstract. Functional magnetic resonance imaging (fMRI) produces low number of samples in high dimensional vector spaces which is hardly adequate for brain decoding tasks. In this study, we propose a combination of autoencoding and temporal convolutional neural network architecture which aims to reduce the feature dimensionality along with improved classification performance. The proposed network learns temporal representations of voxel intensities at each layer of the network by leveraging unlabeled fMRI data with regularized autoencoders. Learned temporal representations capture the temporal regularities of the fMRI data and are observed to be an expressive bank of activation patterns. Then a temporal convolutional neural network with spatial pooling layers reduces the dimensionality of the learned representations. By employing the proposed method, raw input fMRI data is mapped to a low-dimensional feature space where the final classification is conducted. In addition, a simple decorrelated representation approach is proposed for tuning the model hyper-parameters. The proposed method is tested on a ten class recognition memory experiment with nine subjects. Results support the efficiency and potential of the proposed model, compared to the baseline multi-voxel pattern analysis techniques.

1 Introduction

Modeling the relationship between the brain activation patterns and the stimuli is beneficial for understanding the neural code. When the brain activity is recorded during a stimulus, the relationship between the recorded signal and the stimulus category may provide useful information for the underlying cognitive process. Data driven approaches, such as Multi-Voxel Pattern Analysis (MVPA), formulate this relationship as a machine learning task. The problem of predicting the stimulus from the brain recordings, is called brain decoding which has received much attention recently [6]. One of the major difficulties in formulating the brain decoding as a machine learning task is the scarcity of labeled samples. The number of labeled fMRI images over several time points

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-27929-9_3) contains supplementary material, which is available to authorized users.

K.K. Bhatia and H. Lombaert (Eds.): MLMMI 2015, LNCS 9487, pp. 25–34, 2015.
DOI: 10.1007/978-3-319-27929-9_3

reaches at most hundreds, while the dimension of the input space (number of voxels) easily exceeds thousands. In order to improve the classification accuracy and significance, there has been a great deal of effort to either reduce the dimensionality [12] or to employ spatial/temporal structures [14]. As a result, many brain decoding systems rely on cleverly hand-crafted features [5,13,15]. Another issue related to the low number of labeled samples is the labeling procedure. Within the time points acquired during an experiment, only few are assigned to corresponding class labels by considering timing and the type of the stimulus. For example in an event related design, class labels are assigned to time points according to the prior knowledge of the peaks of hemo-dynamic response function [13]. The rest of the (unlabeled) samples are generally discarded from the data sets before any classification pipeline. A typical fMRI experiment for brain decoding is depicted in Fig. 1.

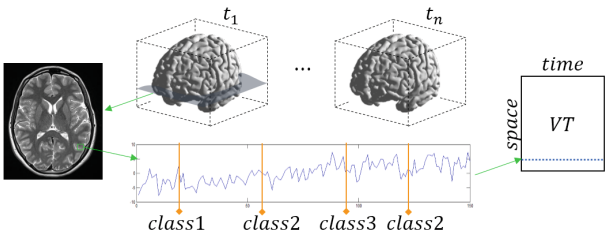


Fig. 1. 4D fMRI data consists of brain volumes across time, some of which are assigned to a class label considering the type and timing of the stimulus. The classification problem is to predict the class of a labeled sample given the rest of the samples. Majority of the samples are unlabeled which are recorded during the inter stimulus side-tasks or resting periods. The whole experimental data is used by forming a data matrix, VT, where the rows correspond to voxel time-series (time axis) and columns correspond to voxels in a 3D volume (space axis) (Color figure online).

Recent improvements in unsupervised feature learning and transfer learning point out the importance of employing unlabeled data for a better classification performance [4], especially, when there is no sufficient amount of labeled data. It is hypothesized that learning the data distribution $p(X)$ for the dataset X from unlabeled samples improves the posteriors $p(Y \mid X)$ of the target Y, when $p(X)$ and $p(Y \mid X)$ share some structure [2]. We experimented this hypothesis by proposing that the discarded (unlabeled) fMRI data may comprise some information and can be exploited in brain decoding. Learning the data distribution $p(X)$ is formulated as learning a set of basis functions that can be used to represent temporal behavior of brain activity. Concretely, by incorporating the unlabeled data, we learn a hierarchy of temporal filters by making use of autoencoders. Learning the posterior of target, $p(Y \mid X)$ is then formulated as a mapping between learned representations and the class labels. Likewise this mapping is modelled by a temporal convolutional neural network (tCNN) whose filters (parameters) pretrained by autoencoders and having additional spatial pooling layers to reduce the dimension of feature space.

By leveraging unlabelled data and the representational power of tCNN, we are able to train comparably better classifiers in terms of classification accuracy and improve generalization performance by reducing the over fitting. The feature spaces are automatically formed by temporal representations and have a reduced dimensionality due to spatial pooling. Although similar approaches are common in computer vision literature [10], there exist only few studies which employ deep learning methods for neuroimaging data [3,7], and to our knowledge the proposed model is the first attempt to use deep learning to learn multiple levels of temporal representations for brain decoding.

2 Unsupervised Learning of Temporal Representations and Convolutional Architecture

In this section, first we explain the structure of the fMRI data and suggest a temporal representation. Then, we introduce a temporal convolutional neural network architecture for brain decoding problem.

Sparse Autoencoders and fMRI Representation. fMRI measurements consist of 3D brain volumes across time $\{t_i\}_{i=1}^{n}$. For each time instant, a 3D brain volume is formed by stacking several 2D slices (scans), see Fig. 1. A task fMRI experiment consists of several runs in which the subjects are presented task specific stimuli at predetermined time instants. Each of these stimuli corresponds to a category and the data acquired at that instant (with a few points of lag) is assigned to the corresponding class label (orange lines in Fig. 1). In this study, the entire experimental data is represented by a voxel×time matrix VT, having n columns and m rows. A column of VT matrix represent the brain volume in terms of m voxels (space) and the rows correspond to the voxel time series of length n.

For classifying individual cognitive states, the labeled samples are separated into training and test sets. Although the columns of VT that do not have a class label are not directly related to the category of stimulus, they may carry significant information about the temporal structure (activation pattern) of the voxel time series. In order to improve the representation power of the voxels, it is crucial to capture and employ the temporal activation pattern as a substitute for raw intensity values. A plausible option is to use unsupervised feature learning methods to learn this temporal aspect. In this study, we utilize the entire VT matrix (except the columns separated for test) to learn the temporal filters. Our aim is to learn a number of HRF-like activation patterns delimited by short time-windows and capture the temporal regularities in the data. For the aim of learning temporal filters in an unsupervised fashion, we employ sparse autoencoders [8].

An autoencoder is a neural network which attempts to reconstruct its input. A 3-layer autoencoder has an input, hidden and output layers, each having several units. Let $x \in \mathbb{R}^{\tau}$ be a given input, the network has an encoder function $f(x;\theta_1)$ which maps x to a hidden representation $h \in \mathbb{R}^{k}$ parametrized by θ_1.

And the decoder function $g(h;\theta_2)$ maps the hidden representations h to the reconstruction of input x which is $\tilde{x} \in \mathbb{R}^\tau$ with learned parameters θ_2. In our problem the inputs x are 1 dimensional patches (temporal-windows) extracted from the rows of VT matrix.

Let $W \in \mathbb{R}^{k\times\tau}$ and $b \in \mathbb{R}^k$ be the weight matrix and biases of the encoder function f such that $\theta_1 = \{W, b\}$ and $W^* \in \mathbb{R}^{\tau\times k}$ and $b \in \mathbb{R}^\tau$ be the weight matrix and biases of the decoder function g with $\theta_2 = \{W^*, b^*\}$,

$$h = f(x;\theta_1) = \sigma(Wx + b), \tag{1}$$

$$\tilde{x} = g(f(x;\theta_1);\theta_2) = \sigma(W^*h + b^*), \tag{2}$$

where σ is an element-wise non-linearity function. Further by enforcing a sparsity constraint (activation around zero) to hidden layer activations via the cost function, auto encoder learns a compact and non-linear representation of its input x. The number of hidden neurons, will be referred as k, is equal to the number of filters (bases) to be learned. The sparse autoencoder having k hidden neurons is trained to minimize reconstruction error using gradient descent by minimizing the cost function $J_{sparse}(\Theta)$ as,

$$J_{sparse}(\Theta) = J_{NN}(\Theta) + \beta J_{\hat{\rho}} + \lambda||\Theta||_2^2, \tag{3}$$

$$J(\Theta) = \frac{1}{2}\sum_i ||\tilde{x}^{(i)} - x^{(i)}||_2^2 + \beta\sum_j^k KL(\rho||\hat{\rho}_j) + \lambda||\Theta||_2^2 \tag{4}$$

where $J_{NN} = \frac{1}{2}\sum_i ||\tilde{x}^{(i)} - x^{(i)}||_2^2$ is the neural network reconstruction term, $\lambda||\Theta||_2^2$ is the L2 regularization and β is the hyper-parameter controlling importance of sparsity in the cost, we use superscript $^{(i)}$ for different examples. The sparsity term $J_{\hat{\rho}}$ is the crucial term in our autoencoder. Let $a_j(x)$ be the activation of hidden unit j given input x and $\hat{\rho}_j = \mathbb{E}[a_j(x)]$ be the expected activation of hidden unit j over the dataset. By constraining $\hat{\rho}_j \approx \rho$ where ρ is the sparsity hyper-parameter, hidden layer activations can be adjusted to be sparse. And to measure the sparsity cost, KL-divergence between average activation of a unit $\hat{\rho}$ and sparsity parameter ρ is calculated as $J_{\hat{\rho}} = \sum_j^{k_1} KL(\rho||\hat{\rho}_j)$.

The optimization of the cost function (3) yields the model parameters $\Theta = \{\theta_1, \theta_2\}$, and rows of the transition weights W of encoder function $f(x;\theta_1)$, constitutes the temporal filters, basis functions to represent the input (see Fig. 2 for sample filters). In our proposed model, we train an autoencoder on a set of randomly selected 1-dimensional patches of size $1 \times \tau_l$ for layer l and use the learned parameters of the encoder as filters in the convolutional layer l. For the first layer, samples are collected from the original VT matrix and for the forthcoming layers, samples are collected from the convolution layer response maps of the previous layer, which will be explained in the next section.

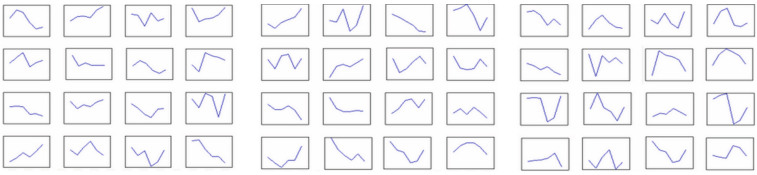


Fig. 2. First level temporal filters learned from unlabelled data for three subjects. Each small box represents a temporal filter and 16 filters for each subject.

1D Convolutional Neural Networks. A convolutional neural network is a feed-forward neural network with local connectivity pattern for hidden units and parameter sharing between inputs of the hidden units [11]. In a convolutional layer each hidden layer has several feature maps and all the hidden units within a feature map share the same parameters. Let us first define a convolutional layer with k feature maps $\{r^i\}_{i=1}^{k}$, as each feature map r^i be 2-dimensional array, having 2-dimensional input $x \in \mathbb{R}^{m \times n}$. Further let W be a filter bank of size $k \times \tau$ and W^i be i^{th} row of W which is a 1-dimensional filter connecting the input to the r^{th} feature map and b^i be the scalar bias for the k^{th} feature map. Given input x, filters and biases W, b and filter size τ, the response at location (m,n) of the feature map r^i is calculated as,

$$r^i(m,n) = \left[\sum_{u=0}^{\tau-1} W^i(\tau - u) x(m, n+u) \right] + b^i, \tag{5}$$

paranthesis indicates indexing of matrix and vectors. With proper zero-padding of the input x each feature map will be of size of its input and a layer constructed of several feature maps obtained with (5) is called a convolutional layer.

Starting from the raw input matrix VT we first collected 50.000 random windows of size τ_1 resulting input matrix of size $x \in \mathbb{R}^{\tau_1 \times 50K}$ where we trained an autoencoder with k_1 hidden units to learn parameters $W_1 \in \mathbb{R}^{k_1 \times \tau_1}$. Rows of W_1 as filters of the first convolutional layer, VT matrix is convolved along the time axis (1D full temporal convolution) with the learned filters using (5) and k_1 number of response matrices are extracted. Note that resulting response matrices are the same size as VT with a full convolution by zero padding. In a CNN, general processing block is constructed by three consecutive operations; convolution, pooling and applying an element-wise non-linearity [9] and these blocks can be repeated consecutively by feeding output of one block to the next one [10]. We construct our temporal CNN with two processing blocks as shown in Fig. 3 with dashed boxes. In order to complete our first block, we determine a spatial pooling function μ, a pooling range in a vicinity δ_1 and a point-wise non-linearity function σ. For layer j, given a feature map r_j^i, pooling function μ and pooling range δ_j, we calculate the pooled response map p_j^i at position (m,n) as follows,

$$p_j^i(m,n) = \mu \left(r_j^i \left(\eta \left(m; \delta_j \right), n \right) \right), \tag{6}$$

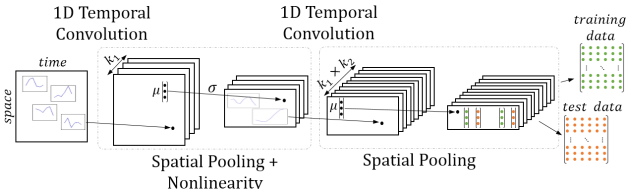


Fig. 3. tCNN with spatial pooling units used for brain decoding. All the filters shown are learned by autoencoders and used for convolution.

where $\eta(p;q)$ is a function that maps voxel at location p to a list of its q nearest neighbor indices including p. Hence μ, takes a list of responses gathered around a voxel at position m as input and returns their maximum. Since consecutive convolution and pooling operations do not break down the neighboring structure of VT matrix, η can be used as many layers are stacked. Considering the capillary structure of the brain and point spread function of fMRI medium, nearby voxels exhibit correlated intensities. Therefore for the choice of μ we employ max function on the columns of feature maps. Note that the columns of the response matrices correspond to spatial domain of fMRI data. As we increase the pooling range without any overlaps between pooling functions, the final dimensionality decreases. We finalize the first processing block by applying σ, which is set to hyperbolic tangent, to all elements of the pooled response matrices.

The second level of our tCNN takes the pooled response matrices as input and repeats the same pipeline. Training an autoencoder for each feature map, each having k_2 hidden units, and applying colvolutions with learned filters by following (5). The only difference for the second block is the number of input matrices, which is k_1 number of m/δ_1 rowed and n columned pooled response matrices. For each of these matrices, a separate k_2 number of temporal filters are learned. Same procedure is followed k_1 times: first collecting τ_2 length time windows from the rows of an input response matrix then training an autoencoder having k_2 hidden units. Each k_2 filters are then convolved with the corresponding first level pooled response matrix. At the end, we obtain $k_1 \times k_2$ number of second level pooled response matrices. Note that second level pooled response matrices have $m/(\delta_1 \times \delta_2)$ rows and n columns, where δ_2 is the pooling range for second processing block. See Supplementary Material (S1) for more details.

After processing raw input data with tCNN, the second block pooled response matrices have the same number of columns n, hence still carries the class label information. Finally at this point, all the second level pooled response matrices are concatenated along their time axis resulting our final feature matrix which has $(m \times k_2)/(\delta_1 \times \delta_2)$ rows and n columns. We then separate the training and test data by extracting corresponding columns of VT matrix, according to the experiment design class labels positions.

3 Experiments and fMRI Data

fMRI recording was conducted during a recognition memory task on nine participants. Each participant is shown a list of words belonging to one of ten categories in the working memory encoding phase. Following a delay period, participant executes a yes/no response indicating the probe word belongs to the current study list. For the decoding task, we select the anterior lateral temporal cortex region having 1024 voxels (m). fMRI data consists of 2400 time points (n) in 8 runs, with 240 labeled samples for the memory encoding phase and 240 labeled samples for the retrieval. The task we seek to accomplish is to predict class labels of the samples in the retrieval phase by using samples in the encoding phase. Measurements recorded in the memory encoding phase are used in training and retrieval samples are used in test. The recording was conducted using a 3T Siemens scanner with a 2 s TR.

Selection of Hyper-parameters. A major bottleneck of designing a deep learning architecture is the large number of hyper-parameters to be tuned [2]. In this study, we propose a heuristic method motivated by the distributed representations and decorrelated features [1]. We assume that the validity of the network can be assured separately at each layer by reducing the redundancy within the learned representations. In other words, we expect that the hidden representations of an autoencoder learns different regularities from the input. The representation power will, then, be increased and diversified as the learned filters are decorrelated. Therefore, we select the least correlated features in a parameter search space for the hyper-parameters of autoencoders $k_1, k_2, \beta, \rho, \lambda$. For each learned filter bank in a layer with a hyper-parameter combination, we calculate the correlation matrix of learned filters. In fully decorrelated filters case, the diagonalized correlation matrix should be close to the identity matrix (all eigenvalues should be equal to one). Therefore, we calculated the L1 distance between the eigenvalues of the correlation matrix of the learned filters to a vector of ones. For example, the hyper-parameters k_1, β and ρ for the first layer autoencoder are calculated as follows,

$$\arg\min_{k_1,\beta,\rho} ||diag(I) - eigs(R_{\theta_1})||_1 \tag{7}$$

where I is the identity matrix and R_{θ_1} is the correlation matrix for filters θ_1. The first layer temporal representations are illustrated in Fig. 2, where we observe several filters that resemble the HRF. Moreover, the filters which directly learn from the unlabelled data are capable of representing several other activation patterns such as linear trends, boxcar, rapid dips and peaks, even some baseline trends.

Testing Procedures and Comparison. In order to test and compare the proposed method with the MVPA methods, we employ the k-nearest neighbor (k-NN) classifier to the feature vectors extracted from the row voxel intensity

values. Three different MVPA approaches are taken into consideration. The first one is the Raw MVPA method where the raw voxel intensity values are directly fed into the classifier. To make a fair comparison, we further employ temporal information in classical MVPA method in two different ways: In the first method, we extract a set of hand-crafted features by convolving the raw intensity values with a double gamma HRF function spanning 6 samples. It is expected that these hand-crafted features capture the temporal activation patterns in the voxel intensity values. This method, called HRF MVPA, trains a k-NN classifier from the response matrix obtained at the output of the convolution. In the second method, which is called Temporal MVPA (T-MVPA), we take a 6 sample time window from the onset of the stimulus and concatenate the subsequent intensity values acquired during that period. This yields a feature vector of dimension 6 for each voxel which is 6 times higher than the other MVPA methods. We also test the single-level temporal convolutional network to monitor the impact of depth on the performance of the classifier.

Experimental results are analysed by considering final feature space dimensions (last column of Table 1) in classifiers by comparing classification accuracies. Overall results are also illustrated in Table 1. For the classical and temporal MVPA approaches (first three rows of Table 1), Raw MVPA approach is taken as baseline method. HRF convolved MVPA model improved baseline up to 8 % as we employ temporal information. However, without adjusting HRF to the regularities in the data, it remains rather hand-crafted. Similarly T-MVPA method use temporal information without any HRF assumption and improves performance compared to other MVPA methods with an increased feature dimensionality. Proposed tCNN architecture is tested with varying depth (single and two layer rows in Table 1) and pooling ranges (δ_1 and δ_2 columns). A single layer (shallow) temporal feature learning and convolution architecture with varying pooling range between 2 to 16 yields a feature space with dimensions from 8192 to 1024. For all nine subjects, single layer architecture outperforms classical and temporal MVPA methods substantially and achieving performance up to 60 %s. By appropriate pooling, the feature dimensions of the single layer architecture retracted down to 1024 where we still observe 20 % improvement.

In order to analyse the impact of the depth and further reduce feature dimensionality, two layer architecture is tested with varying pooling ranges in the second level. The proposed model in two layers reduces feature dimensions down to 256 where we still observe better performance in 8 subjects compared to the single layer architecture with lowest dimensions. We did not observe any performance improvements by increasing depth more than two, as model gets more complex and starts to overfit rapidly. The results suggest that employing temporal structures with appropriate pooling for brain decoding gives rise to better classification accuracies. Convolutional models pre-trained with a slightly larger amount of unlabelled data are appropriate candidates to employ temporal information in an efficient way. Furthermore, increasing depth of such temporal convolutional architectures, makes it possible to reduce feature dimensionality, and the learned filters become more abstract and non-linear, resulting a better

Table 1. Comparison of classification accuracies for the proposed method.

Method	δ_1	δ_2	Sub1	Sub2	Sub3	Sub4	Sub5	Sub6	Sub7	Sub8	Sub9	Dim
Raw MVPA	-	-	28.9	41.5	43.1	45.5	42.3	31.9	34.0	40.6	41.4	1024
HRF MVPA	-	-	32.3	44.4	42.3	53.8	36.1	41.1	44.5	46.5	48.2	1024
T-MVPA	-	-	41.9	52.4	52.0	53.9	45.3	44.0	44.9	51.1	52.5	6144
Single layer	2	-	54.0	65.7	65.3	67.8	66.7	64.1	66.1	65.7	69.5	8192
Single layer	4	-	56.1	64.5	63.6	69.5	66.2	65.3	66.1	66.5	68.6	4096
Single layer	8	-	55.7	65.3	62.8	69.5	64.0	64.9	66.1	67.4	70.0	2028
Single layer	16	-	51.9	64.0	63.2	68.6	62.3	63.3	64.9	67.3	67.0	1024
Two layers	4	2	69.0	**71.6**	72.9	72.4	**72.8**	**71.2**	**72.3**	70.3	71.2	8142
Two layers	16	1	68.6	69.1	71.2	72.4	69.1	68.6	60.7	70.3	72.8	4096
Two layers	16	2	69.5	68.6	71.6	**73.2**	70.0	69.9	66.1	70.3	**72.8**	2048
Two layers	16	4	**69.9**	69.5	**73.3**	72.8	71.2	70.0	67.0	69.5	71.6	1024
Two layers	16	8	68.6	67.4	72.0	70.8	71.1	68.7	67.0	**70.3**	70.8	512
Two layers	16	16	63.6	66.5	69.9	70.7	68.3	66.6	64.4	68.2	66.5	256

classification performance. The fewer number of training samples may result in greater variability of the estimation. Therefore we conducted simple, yet diagnostic tests to determine how precise the proposed method performs where details can be found in Supplementary Material (S2).

4 Conclusion

In this study, we propose a novel approach for brain decoding on fMRI data using unsupervised feature learning and convolutional neural networks. By leveraging unlabelled data and employing multi-layer temporal CNNs, we learn multiple layers of temporal filters which represent the activation patterns of voxels under experimental conditions. By making use of deep temporal representations, we train comparatively better brain decoding models in terms of classification performance. This method suggests a shift from the conventional MVPA approaches which rely on the hand-crafted features, to learned feature representations of deep architectures. As an evidence of the power of proposed model, we conducted a recognition memory experiment on 9 subjects and observed significant performance improvements. The proposed model has potential to further improvements by incorporating spatial structures with spatial convolution and pooling, or learning spatio-temporal filters all-together which is left as a future work. Also, hyper-parameter selection and preventing overfitting is partially and naively handled in this study, which should be further analysed in details.

References

1. Bengio, Y., Bergstra, J.S.: Slow, decorrelated features for pretraining complex cell-like networks. In: Advances in Neural Information Processing Systems, pp. 99–107 (2009)

2. Bengio, Y., Courville, A., Vincent, P.: Representation learning: a review and new perspectives. IEEE Trans. Pattern Anal. Mach. Intell. **35**(8), 1798–1828 (2013)
3. Cadieu, C.F., Hong, H., Yamins, D.L.K., Pinto, N., Ardila, D., Solomon, E.A., Majaj, N.J., DiCarlo, J.J.: Deep neural networks rival the representation of primate it cortex for core visual object recognition. PLoS Comput. Biol. **10**(12), e1003963 (2014)
4. Erhan, D., Bengio, Y., Courville, A., Manzagol, P.A., Vincent, P., Bengio, S.: Why does unsupervised pre-training help deep learning? J. Mach. Learn. Res. **11**, 625–660 (2010)
5. Hausfeld, L., Valente, G., Formisano, E.: Multiclass fmri data decoding and visualization using supervised self-organizing maps. NeuroImage **96**, 54–66 (2014)
6. Haxby, J.V., Connolly, A.C., Guntupalli, J.S.: Decoding neural representational spaces using multivariate pattern analysis. Ann. Rev. Neurosci. **37**(1), 435–456 (2014)
7. Hjelm, R.D., Calhoun, V.D., Salakhutdinov, R., Allen, E.A., Adali, T., Plis, S.M.: Restricted boltzmann machines for neuroimaging: an application in identifying intrinsic networks. NeuroImage **96**, 245–260 (2014)
8. Kavukcuoglu, K., Ranzato, M., Fergus, R., LeCun, Y.: Learning invariant features through topographic filter maps. In: IEEE CVPR, pp. 1605–1612 (2009)
9. Krizhevsky, A., Sutskever, I., Hinton, G.E.: Imagenet classification with deep convolutional neural networks. In: Advances in Neural Information Processing Systems (2012)
10. Le, Q.V., Zou, W.Y., Yeung, S.Y., Ng, A.: Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In: IEEE CVPR, pp. 3361–3368 (2011)
11. LeCun, Y., Bottou, L., Bengio, Y., Haffner, P.: Gradient-based learning applied to document recognition. Proc. IEEE **86**(11), 2278–2324 (1998)
12. Mwangi, B., Tian, T., Soares, J.: A review of feature reduction techniques in neuroimaging. Neuroinformatics **12**(2), 229–244 (2014)
13. Norman, K.A., Polyn, S.M., Detre, G.J., Haxby, J.V.: Beyond mind-reading: multi-voxel pattern analysis of fMRI data. Trends Cogn. Sci. **10**(9), 424–430 (2006)
14. Pereira, F., Botvinick, M.: Information mapping with pattern classifiers: a comparative study. Neuroimage **56**(2), 476–496 (2011)
15. Shirer, W.R., Ryali, S., Rykhlevskaia, E., Menon, V., Greicius, M.D.: Decoding subject-driven cognitive states with whole-brain connectivity patterns. Cereb. Cortex **22**(1), 158–165 (2012). (New York, N.Y.: 1991)

Modelling Non-stationary and Non-separable Spatio-Temporal Changes in Neurodegeneration via Gaussian Process Convolution

Marco Lorenzi[1(✉)], Gabriel Ziegler[2], Daniel C. Alexander[3], and Sebastien Ourselin[1]

[1] Translational Imaging Group, CMIC, UCL, London, UK
m.lorenzi@ucl.ac.uk
[2] Wellcome Trust Centre for Neuroimaging, UCL, London, UK
[3] Centre for Medical Image Computing (CMIC), UCL, London, UK

Abstract. Modelling longitudinal changes in organs is fundamental for the understanding of biological and pathological processes. Most of the previous works on spatio-temporal modelling of image time series relies on the assumption of stationarity of the local spatial correlation, and on the separability between spatial and temporal processes. These assumptions are often made in order to lead to computationally tractable approaches to longitudinal modelling, but inevitably lead to an oversimplification of the complex spatial and temporal dynamics underlying the biological processes. In this work we propose a novel spatio-temporal generative model of time series of images based on kernel convolutions of a white noise Gaussian process. The proposed model is parameterised by a sparse set of control points independently identified by specific spatial and temporal parameters. This formulation is highly flexible and can naturally account for spatially and temporally varying dynamics of changes. We demonstrate a preliminary application of our non-parametric method on the modelling of within-subject structural changes in the context of longitudinal analysis in Alzheimer's disease. In particular we show that our method provides an accurate description of the pathological evolution of the brain, while showing high flexibility in modelling and predicting region-specific non-linearity due to accelerated structural decline in dementia.

1 Introduction

Modelling longitudinal changes in organs is fundamental for the understanding of biological and pathological processes. For instance the development of a spatio-temporal model of disease progression in Alzheimer's disease (AD) from time

S. Ourselin—Data used in preparation of this article were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database (www.loni.ucla.edu/ADNI). As such, the investigators within the ADNI contributed to the design and implementation of ADNI and/or provided data but did not participate in analysis or writing of this report. A complete listing of ADNI investigators can be found at: www.loni.ucla.edu/ADNI/Collaboration/ADNI_Authorship_list.pdf.

K.K. Bhatia and H. Lombaert (Eds.): MLMMI 2015, LNCS 9487, pp. 35–44, 2015.
DOI: 10.1007/978-3-319-27929-9_4

series of magnetic resonance images (MRIs) would be highly valuable for the fundamental understanding of the disease process, for diagnostic purposes and individual predictions, and for testing the efficacy of disease modifying drugs in clinical trials.

The consistent modelling and prediction of spatio-temporal changes in longitudinal MRI is still an important challenge from both methodological and computational perspectives. In fact, flexible modelling instruments are required in order to robustly capture meaningful pathological accelerations specific to sensitive brain regions. Moreover, since a biological model of local brain changes is often unknown, it is important to develop optimal models in terms of statistical complexity. Notably, the spatial dimensionality of MRI time series prevents the straightforward implementation of classical multivariate statistical modelling techniques and often leads to computationally intractable solutions.

We can identify two main approaches to spatio-temporal modelling of image time-series in computational anatomy. The first one is based on non-linear image registration, describing signal differences between images as local spatial transformations [1–4]. In non-linear registration the spatial changes are usually modelled at a fixed spatial scale defined by the regularization energy at which the transformation is optimized. The temporal modeling usually relies on the definition of a specific model of temporal evolution, which is identified either by fitting parametric progression models on geometric features of the transformation, or by choosing an opportune metric in the space of transformations to characterize specific evolution models in the image space. The second one, usually identified as voxel-based-morphometry (VBM), is based on voxel-by-voxel modelling based on parametric [5], or non-parametric regression frameworks [6]. Models are usually independently fitted for each voxel, and local correlation is usually imposed by applying Gaussian convolution of the images with some apriori kernel size.

The majority of the above mentioned approaches rely on important assumptions concerning the spatial and temporal processes. In fact, by either choosing a global regularization energy in image registration, or a global smoothing parameter in VBM, we usually impose local *stationary* correlation models for the spatial changes. Even though this assumption is often necessary to lead to computationally tractable approaches, it inevitably leads to an oversimplification of the complex spatial properties of the images, for example concerning regionally varying smoothness, and image boundaries.

At the same time, by fitting global longitudinal models, either defined by the registration metric, or by a fixed statistical model complexity, we assume that spatial and temporal processes are *separable*, i.e. that the properties of the temporal variation (for instance following a quadratic or linear behaviour) is independent from the spatial locations. As before, this assumption often leads to simplistic modeling solution, as the progression of the temporal changes in organs is generally highly variable across spatial regions.

Non-parametric Gaussian process (GP) models have emerged as a flexible and elegant Bayesian approach for prediction of continuous and binary variables in manifold applications [7]. Recently, GPs were successfully introduced to the field

of neuroimaging, e.g. in the context of single-case inference in aging [6]. Moreover, it was recently introduced in [8] a generative framework for the modelling of image time series based on Gaussian process regression. This approach was however completely based on the assumption about stationarity and separability of spatial and temporal processes.

In this work we propose a generative model of image time series based on Gaussian processes (GPs), which is characterized by a covariance structure parameterized by a sparse set of control points defined in space and time. Since each control point is governed by specific spatial and temporal parameters, the proposed model is highly flexible and can naturally account for spatially and temporally varying signal changes. The proposed model thus overcomes many limitations of previous spatio-temporal modelling approaches.

The paper is organized as follows. In Sect. 2 we propose our generative model of longitudinal changes parameterized by a sparse set of control points, and we subsequently provide details about parameter optimization and prediction. In Sect. 3 we provide a preliminary application of our non-parametric method on the modelling of within-subject structural changes in the context of longitudinal analysis in Alzheimer's disease. In particular we show that our method provides an accurate description of the pathological evolution of the brain, while showing high flexibility in modelling and predicting region-specific non-linearity due to accelerated structural decline in dementia.

2 Generative Model of Spatial Data Through Gaussian Process Convolution

Let u and t be respectively the spatial and temporal coordinates. Given an image time series $y(s)$, $s = (u, t)$, we assume a generative model for the spatio-temporal variations:

$$y(s) = z(s) + \epsilon, \tag{1}$$

where ϵ Gaussian distributed spatial noise $\epsilon \sim \mathcal{N}(0, \sigma_\epsilon^2)$, and where z is a (zero-mean) Gaussian process (GP), identified by the associated covariance form Σ. Following the idea introduced in [9], we model $z(s)$ as the convolution of a white noise process $x(s) \sim \mathcal{GP}(0, \sigma_x^2 Id)$, with a given kernel function k. More specifically, the spatial process $z(s)$ is identified by a sparse set of control points defined in space and time, $\{w_j = (u_j, t_j))\}_{j=1}^{N_w}$, and associated parameters θ_j:

$$z(s) = \sum_{j=1}^{N_w} x(w_j) k(s - w_j | \theta_j). \tag{2}$$

Under these modelling assumptions, the generative model (1) assumes the form:

$$y = Kx + \epsilon, \tag{3}$$

where K is the matrix of the spatial coefficients associated to the control points $K_{s,w_j} = k(s - w_j|\theta_j)$. The image time series y is therefore a realization of the following process:

$$y \sim \mathcal{GP}(0, \Sigma), \quad \text{with } \Sigma = \sigma_x^2 K K^T + \sigma_\epsilon^2 Id. \tag{4}$$

The model (4) is completely identified by the measurement noise σ_ϵ, by the white process parameter σ_x, and by the control points w_j with associated parameteres θ_j.

We note that the size of the covariance matrix Σ is $N \times N$, where $N = N_u N_t$, and N_u is the number of voxels, and N_t is the number of temporal observations. For this reason the naive approach to the modelling of (4) can easily lead to prohibitive problems in term of storage and computations. In the following section we show that the proposed kernel parameterization leads to computationally tractable inference schemes whose complexity depends on the number of basis functions.

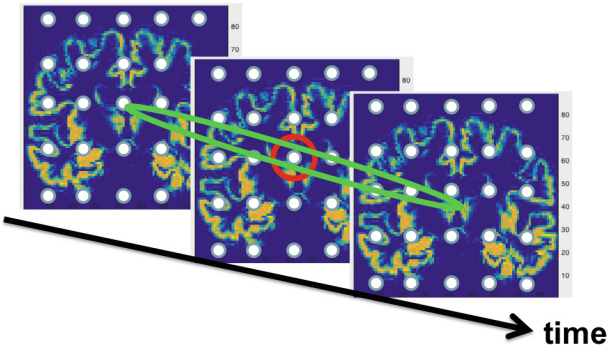


Fig. 1. The parameters associated to each control point of the grid completely identify the local spatial and temporal covariance (exemplified by resp. red and green neighborhoods) of the spatio-temporal process (Color figure online).

2.1 Efficient Inference in Gaussian Process Convolution Models

The GP-based generative model with kernel structure outlined in this work provides a powerful and extremely flexible framework for prediction an inference in image time series. Let $\boldsymbol{\theta} = \{\sigma_x, \sigma_\epsilon, (\theta_j)_{j=1}^{N_w}\}$ be the set of parameters of the model (4). In the following sections we provide the main results concerning the marginal likelihood computation, the hyper-parameter optimization and the posterior prediction.

2.2 Log-Marginal Likelihood

The log-marginal likelihood of model (4) is:

$$\log \mathcal{L}(\boldsymbol{\theta}) = -\tfrac{n}{2}\log(2\pi) - \tfrac{1}{2}\log|\Sigma(\boldsymbol{\theta})| - \tfrac{1}{2}y^T\Sigma(\boldsymbol{\theta})^{-1}y. \tag{5}$$

In particular, the determinant and matrix inverse terms can be efficiently computed by using well known matrix algebra properties:

$$|\Sigma(\boldsymbol{\theta})| = |\sigma_x^2 K^T K + \sigma_\epsilon^2 Id_{N_w}| \tag{6}$$

$$\Sigma(\boldsymbol{\theta})^{-1} = \frac{1}{\sigma_\epsilon^2} Id_N - \frac{1}{\sigma_\epsilon^4} K \left(\frac{1}{\sigma_x^2} Id_{N_w} + \frac{1}{\sigma_x^2} K^T K \right)^{-1} K^T. \tag{7}$$

We note that in this form both inverse and determinant operations are performed on matrices of size N_w, which is magnitude smaller than N.

2.3 Hyperparameter Optimization

The derivative of the log-marginal likelihood (5) with respect to the model parameters $\boldsymbol{\theta}$ is:

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} \log \mathcal{L} = -\tfrac{1}{2} Tr \left(\Sigma(\boldsymbol{\theta})^{-1} \tfrac{\mathrm{d}\Sigma(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} \right) - y^T \Sigma(\boldsymbol{\theta})^{-1} \tfrac{\mathrm{d}\Sigma(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} \Sigma(\boldsymbol{\theta})^{-1} y \tag{8}$$

It can be shown that formula (8) can be efficiently computed with respect to each model parameters. For instance, the gradient with respect to the noise parameter can be expressed in the form:

$$\frac{\mathrm{d}}{\mathrm{d}\sigma_\epsilon^2} \log \mathcal{L} = \frac{1}{2} [-N + \frac{1}{\sigma_\epsilon^2} Tr(K^T K A) + \tag{9}$$

$$y^T \left(\frac{1}{\sigma_\epsilon^2} Id_N - \frac{2}{\sigma_\epsilon^4} KAK^T + \frac{1}{\sigma_\epsilon^6} KAK^T KAK^T \right) y], \tag{10}$$

where

$$A = \left(\frac{1}{\sigma_x^2} Id_{N_w} + \frac{1}{\sigma_\epsilon^2} K^T K \right)^{-1}.$$

We note that, as for the computation of the marginal likelihood, the above term can be efficiently decomposed in the more convenient product of matrices of lower dimension, thus leading to computationally tractable solutions. In particular, the computation and allocation of matrices of size $N \times N$ is never required. The explicit derivation of the other model parameters is provided in the Appendix.

2.4 Prediction

The proposed generative model allows us to consider the predictive distributions of the latent spatio-temporal process at any testing locations u^* and timepoints t^*.

Given image time series $I(u, t)$, we now aim at predicting the image I^* at $N^* \times N_T^*$ testing coordinates $\{u^*, t^*\}$. Let us define $\Sigma_{I,I^*} = \Sigma(u, t, u^*, t^*)$

the cross-covariance matrix of training and testing data, and $\Sigma_{I^*,I^*} = \Sigma(u^*, t^*, u^*, t^*)$ the covariance evaluated on the new coordinates. The joint GP model of training and testing data is:

$$\begin{pmatrix} I(u,t) \\ I^*(u^*,t^*) \end{pmatrix} \sim \mathcal{N}\left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma + \sigma^2 Id & \Sigma_{I,I^*} \\ \Sigma_{I^*,I} & \Sigma_{I^*,I^*} + \sigma^2 Id \end{pmatrix} \right] , \tag{11}$$

and it can be easily shown that the posterior distribution of I^* conditioned on the observed time series I and parameters $\boldsymbol{\theta}$ is [7]:

$$\begin{aligned} I^*|I, \{u^*, t^*\}, \boldsymbol{\theta} \sim & \mathcal{N}\Big(\mu^*, \Sigma^*\Big), \text{ where } \mu^* = \Sigma_{I,I^*}\Sigma^{-1}I \\ & \text{and } \Sigma^* = \Sigma_{I^*,I^*} - \Sigma_{I,I^*}\Sigma^{-1}\Sigma_{I^*,I} + \sigma^2 Id \ . \end{aligned} \tag{12}$$

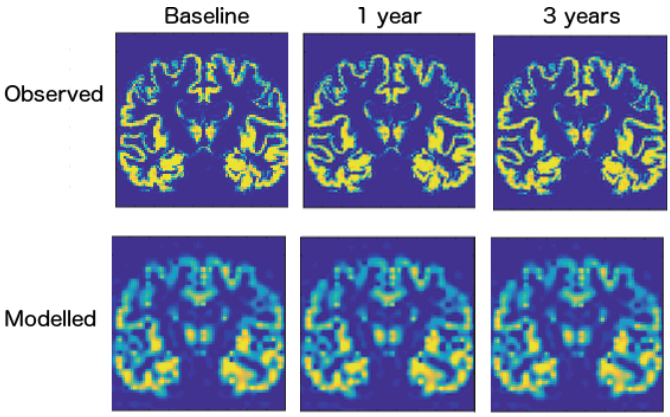


Fig. 2. Observed and modelled image time series by using a grid of 25×25 control points. The predicted progression provides a probabilistic description of the observed data at the grid resolution.

3 Application: Longitudinal Brain Changes in Alzheimer's Disease

In this section we show an application of the proposed generative model to the analysis of the individual longitudinal brain changes observable in image time series. We consider the model outlined in Eq. (3), with kernel function k associated to the control points $\{w_j\}_{j=1}^{N_w}$ identified by independent spatial and temporal length-scale parameters $\theta_j = \{\theta_j^u, \theta_j^t\}$:

$$k(s - w_j|\theta_j) = \exp(-|u - u_j|^2/\theta_j^u)\exp(-|t - t_j|^2/\theta_j^t). \tag{13}$$

With the proposed parameterization, the spatio-temporal process (3) is completely characterized by the sparse set of spatial and temporal parameters associated to the set of control points (Fig. 1). As we shall see in the following experiment, these parameters describe the spatial and temporal complexity of the underlying spatio-temporal signal, and thus they identify the non-stationary and non-separable model of the observed image time series.

3.1 Data Analysis and Results

We selected a patient affected by mild cognitive impairment for which 6 images were available, corresponding to observational time of respectively baseline, 6 months, 1, 1.5, 2 an 3 years.

The images were processed according to established procedures consisting of joint bias correction, tissue segmentation, alignment to the within-subject average anatomy, and non-linear normalization to a group-wise anatomical reference [10]. The final image size was of 100^3 cubic voxels with isotropic resolution of 1.5 mm.

Figure 2 shows an application of the proposed approach on the modeling of the coronal slice including temporal regions, by using a grid of 25×25 basis functions. We note that the predicted progression provides a description of the observed data at the grid resolution.

The fitted model parameters are shown in Fig. 3, left. It is interesting to note that they provide a description of the spatial and temporal complexity of the observed time series. Indeed, the spatial complexity is higher (decreased spatial length-scale parameter) in the cortical areas, while the temporal complexity is higher in the temporal regions (decreased temporal length-scale). We also note that the model variability is zero outside the brain areas (Fig. 3, right).

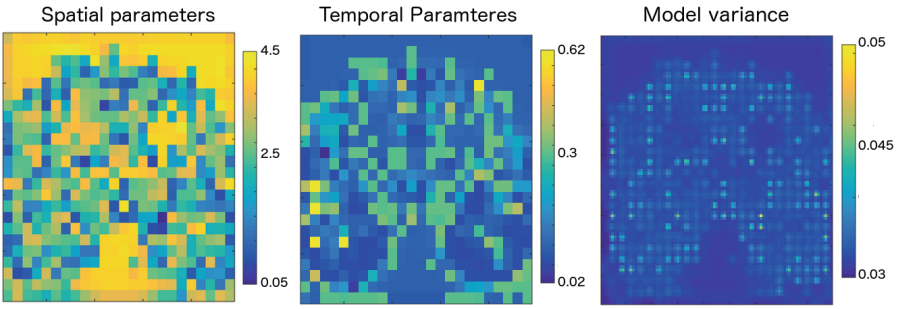


Fig. 3. Fitted model parameters and model variance. We note that the model is capable to adjust the parameters to the spatial and temporal complexity of the data. In particular, the spatial complexity is higher (decreased spatial length-scale parameter) in the cortical areas, while the temporal complexity is higher in the temporal regions (decreased temporal length-scale). We also note that the model variability is zero outside the brain areas.

The accuracy of the proposed approach in modelling the longitudinal changes is shown in Fig. 4, where we show the average longitudinal changes measured in respectively temporal areas and thalami, two regions which are characterized by different temporal complexity (Fig. 3): the temporal length-scale parameters of the temporal region are low (thus denoting high temporal complexity of this area), while the ones of the thalami are associated to higher length-scale (low temporal complexity). Indeed, the average progressions shown in Fig. 4 show an almost constant progression for the thalami, while the temporal area has an accelerated atrophy process.

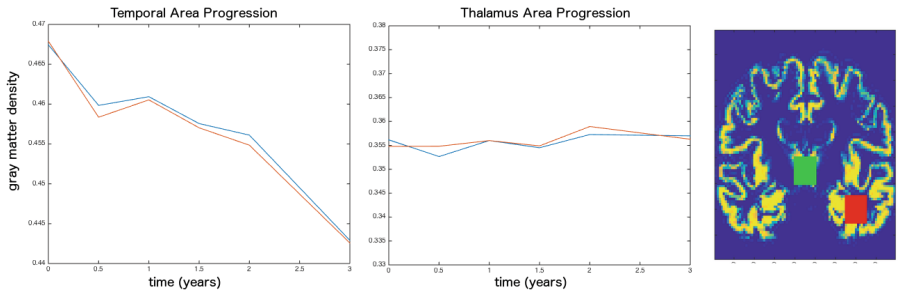


Fig. 4. Left. Modelled (red) and observed (blue) atrophy progressions. The model provides accurate fit, and shows that the temporal areas have higher temporal complexity than the thalami, caused by the process of atrophy acceleration. Right. Reference areas enclosing temporal region (red), and thalami (green) (Color figure online).

4 Conclusions

In this work we proposed a novel probabilistic approach to the modelling of non-stationary, non-separable spatio-temporal processes, by means of kernel convolutions of a white noise Gaussian process. The proposed approach is inspired by the literature on non-stationary models of spatio-temporal changes [11], with particular focus on the study of non-stationary covariance structures [12]. The experimental results show that the proposed modelling method leads to an accurate fit to the observed image time series, while at the same time providing a rich description of the spatio-temporal dynamics of the data, encoded by the learned spatial and temporal parameters. Further extensions of the proposed work will aim at improving the computationally efficiency of the inferential process, in order to scale to the modelling of high-dimensional time series of 3D images with several time points. Finally, the present work will be in the future validated by assessment of the predictive accuracy on large studies, and on different clinical cases, such as asymptomatic or prodromal AD cases, which are characterized by subtle and potentially non-linear dynamics of brain changes non-uniformly localized in the brain.

Appendix

We explicit here the derivatives of the log-marginal likelihood (5) with respect to the model parameters $\boldsymbol{\theta}$:

$$\frac{\mathrm{d}}{\mathrm{d}\boldsymbol{\theta}} \log \mathcal{L} = -\tfrac{1}{2} Tr\left(\Sigma(\boldsymbol{\theta})^{-1} \tfrac{\mathrm{d}\Sigma(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}}\right) - y^T \Sigma(\boldsymbol{\theta})^{-1} \tfrac{\mathrm{d}\Sigma(\boldsymbol{\theta})}{\mathrm{d}\boldsymbol{\theta}} \Sigma(\boldsymbol{\theta})^{-1} y. \tag{14}$$

With the following simplification

$$A = \left(\frac{1}{\sigma_x^2} Id_{N_w} + \frac{1}{\sigma_\epsilon^2} K^T K\right)^{-1}.$$

the derivative are as follows:

– Noise parameter.

$$\frac{\mathrm{d}}{\mathrm{d}\sigma_\epsilon^2} \log \mathcal{L} = \frac{1}{2}[-N + \frac{1}{\sigma_\epsilon^2} Tr(K^T K A) \tag{15}$$

$$+ y^T \left(\frac{1}{\sigma_\epsilon^2} Id_N - \frac{2}{\sigma_\epsilon^4} KAK^T + \frac{1}{\sigma_\epsilon^6} KAK^T KAK^T\right) y]. \tag{16}$$

– Amplitude parameter.

$$\frac{\mathrm{d}}{\mathrm{d}\sigma_x^2} \log \mathcal{L} = -\frac{1}{2} Tr(\frac{\sigma_\epsilon^2}{\sigma_x^2} K^T K) - Tr(\frac{\sigma_x^2}{\sigma_\epsilon^4} K^T KAK^T K) \tag{17}$$

$$+ \frac{\sigma_x^2}{2} \left(\frac{1}{\sigma_\epsilon^2} y^T K - \frac{1}{\sigma_\epsilon^4} y^T KAK^T K\right) \left(\frac{1}{\sigma_\epsilon^2} K^T y - \frac{1}{\sigma_\epsilon^4} K^T KAK^T y\right) \tag{18}$$

– Control points parameters.

$$\frac{\mathrm{d}}{\mathrm{d}\theta_j} \log \mathcal{L} = -\frac{\sigma_x^2}{\sigma_\epsilon^2} Tr(\frac{\mathrm{d}K}{\mathrm{d}\theta_j} K^T) + \frac{\sigma_x^2}{\sigma_\epsilon^4} Tr(K^T KAK^T \frac{\mathrm{d}K}{\mathrm{d}\theta_j}) \tag{19}$$

$$- 2\frac{\sigma_x^2}{\sigma_\epsilon^4} (y^T K \frac{\mathrm{d}K}{\mathrm{d}\theta_j} y) \tag{20}$$

$$+ 2\frac{\sigma_x^2}{\sigma_\epsilon^6} (y^T KAK^T \frac{\mathrm{d}K}{\mathrm{d}\theta_j} K^T y + y^T KAK^T K \frac{\mathrm{d}K}{\mathrm{d}\theta_j} y) \tag{21}$$

$$- 2\frac{\sigma_x^2}{\sigma_\epsilon^8} (y^T KAK^T \frac{\mathrm{d}K}{\mathrm{d}\theta_j} K^T KAK^T y). \tag{22}$$

References

1. Davis, B.C., Fletcher, P.T., Bullitt, E., Joshi, S.C.: Population shape regression from random design data. Int. J. Comput. Vis. **90**(2), 255–266 (2010)

2. Ashburner, J., Ridgway, G.: Symmetric diffeomorphic modeling of longitudinal structural MRI. Front. Neurosci. **6**(197) (2013)
3. Niethammer, M., Huang, Y., Vialard, F.-X.: Geodesic regression for image time-series. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part II. LNCS, vol. 6892, pp. 655–662. Springer, Heidelberg (2011)
4. Lorenzi, M., Ayache, N., Frisoni, G.B., Pennec, X.: Mapping the effects of Aβ 1–42 levels on the longitudinal changes in healthy aging: hierarchical modeling based on stationary velocity fields. In: Fichtinger, G., Martel, A., Peters, T. (eds.) MICCAI 2011, Part II. LNCS, vol. 6892, pp. 663–670. Springer, Heidelberg (2011)
5. Friston, K.J., Holmes, A., Worsley, K.J.: Statistical parametric maps in functional imaging: a general linear approach. Hum. Brain Mapp. **2**, 189–210 (1995)
6. Ziegler, G., Ridgway, G.R., Dahnke, R., Gaser, C.: Individualized Gaussian process-based prediction and detection of local and global gray matter abnormalities in elderly subjects. NeuroImage **97**, 333–348 (2014)
7. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. The MIT Press, Cambridge (2005)
8. Lorenzi, M., Ziegler, G., Alexander, D.C., Ourselin, S.: Efficient Gaussian process-based modelling and prediction of image time series. In: Ourselin, S., Alexander, D.C., Westin, C.-F., Cardoso, M.J. (eds.) IPMI 2015. LNCS, vol. 9123, pp. 626–637. Springer, Heidelberg (2015)
9. Higdon, D.: Space and space-time modeling using process convolutions. In: Anderson, C.W., Barnett, V., Chatwin, P.C., El-Shaarawi, A.H. (eds.) Quantitative Methods for Current Environmental Issues, pp. 37–56. Springer, London (2002)
10. Ashburner, J., Friston, K.: Unified segmentation. NeuroImage **26**, 839–851 (2005)
11. DiMatteo, I., Genovese, C.: Bayesian curve-fitting with free-knot splines. Biometrika **88**, 1055–1071 (2002)
12. Paciorek, C., Schervish, M.: Nonstationary covariance functions for Gaussian process regression. Adv. Neural Inf. Proc. Sys. **16**, 273–280 (2004)

Improving MRI Brain Image Classification with Anatomical Regional Kernels

Jonathan Young[1(✉)], Alex Mendelson[1], M. Jorge Cardoso[1], Marc Modat[1], John Ashburner[2], and Sebastien Ourselin[1,3]

[1] Centre for Medical Image Computing, University College London, London, UK
jonathan.young@ucl.ac.uk
[2] Wellcome Trust Centre for Neuroimaging, Institute of Neurology, University College London, London, UK
[3] Dementia Research Centre, Institute of Neurology, University College London, London, UK

Abstract. Classification of brain images is frequently done using kernel based methods, such as the support vector machine. These lend themselves to improvement via multiple kernel learning, where a number of different kernels are linearly combined to integrate different sources of information and increase accuracy. Previous applications made use of a small number of kernels representing different image modalities or kernel functions. Here, the kernels instead represent 83 anatomically meaningful brain regions. To find the optimal combination of kernels and perform classification, we use a Gaussian Process framework to infer the maximum likelihood weights. The resulting formulation successfully combines voxel level features with prior anatomical knowledge. This gives an improvement in classification accuracy of MRI images of Alzheimer's disease patients and healthy controls from the ADNI database to almost 88 %, compared to less than 86 % using a single kernel representing the whole brain. Moreover, interpretability of the classifier is also improved, as the optimal kernel weights are sparse and give an indication of the importance of each brain region in separating the two groups.

Keywords: Gaussian processes · Classification · Multi-kernel learning · MRI · Alzheimer's disease · Interpretability

1 Introduction

Machine learning methods have become increasingly common in the analysis of brain image data, both for computer aided diagnosis (CAD) of disease and in a more exploratory fashion to discover biomarkers that can be informative about disease processes. For Alzheimer's disease (AD), grey matter (GM) density maps obtained from structural MRI images are used as sources of data in the classification. However the actual features derived from the image can take two forms: as the intensities of MRI voxels themselves [1], or as aggregations of all GM voxels within different anatomical regions. The regions can be defined by

K.K. Bhatia and H. Lombaert (Eds.): MLMMI 2015, LNCS 9487, pp. 45–53, 2015.
DOI: 10.1007/978-3-319-27929-9_5

an atlas [2] or can themselves be generated from voxel level data [3]. There is a trade-off between these methods. Regional level features reduce the data dimensionality and can introduce prior information relevant to the classification problem, but also eliminate fine detail that may be informative about disease state. Voxel level data can introduce noise by including uninformative brain regions and results in a very high dimensional problem. The different feature extraction methods are compared and discussed in depth in [4].

Our proposed method combines the strengths of these two approaches. It uses both voxel level features and atlas derived regions, and automatically gives less weight to voxels within less relevant regions. This is done using multiple kernel learning (MKL), a method that can be applied to any kernel based classifier, such as the support vector machine (SVM) or Gaussian Process (GP). These use a linear combination of kernels, where the kernels can be derived from different data modalities [2,5] or kernel functions [6]. Conversely, in our approach each kernel represents the voxel level data *within a different anatomical region* to produce anatomical regional kernels (ARKs). This takes a similar approach to [7,8] presented a related method using hierarchical groups of regional features. Although the work was developed from our previous use of MKL, and is presented as a specific case of MKL, it is related to other families of methods. Specifically, it can be seen as a way to incorporate explicit spatial regularisation into the classifier. A number of other methods have been developed to do this specifically for three dimensional medical image data. Spatial smoothness and sparsity can be enforced with a joint $\ell 1$ and total variation penalty [9]. Alternative a smoothness penalty is derived from the image voxel neighbourhood structure, which can be built into a kernel function for use with an SVM or other kernel method [10] or used directly as a term in the objective function [11].

Our method and [11] can also both be interpreted as a variant of automatic relevance determination (ARD) [12,13], a Bayesian method of automatic feature selection. Our method, however, operates at the regional level in the kernel space, rather than at the voxel level in the input space. Our approach builds on the existence of a brain atlas in a custom groupwise template. We explain how this was achieved, and how MKL is performed within a GP framework.

We apply this method to a large population of AD and control subjects from the ADNI study. In terms of classification accuracy, our method outperforms a single kernel with voxel level features by a susbtantial margin, and a single kernel with regional features by a smaller amount. We also introduce a new method to assess the quality of a classifier that exploit the probabilistic predictions made by GPs. Finally, we show that the optimal kernel weights in the MKL formulation are informative about which regions are affected by AD.

2 Materials and Methods

2.1 Image and Biomarker Data

All data were obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database[1]. The MRI images were T1 weighted structural scans from

[1] http://adni.loni.ucla.edu/.

a mixture of 1.5T and 3T scanners. All were subjected to quality control and automatically corrected for spatial distortion caused by gradient nonlinearity and B1 field inhomogeneity and downloaded from the ADNI database. Subjects were classified as healthy control (HC), AD or mild cognitive impairment by neuropsychological and clinical testing at the time of the baseline scan, and only HC and AD subjects were used. For the classification experiments, a further quality control step was taken which removed 16 subjects with registration errors, leaving a final total of 627 subjects. Their demographics are given in Table 1.

Table 1. Subject groups and demographics

Disease status	Number	1.5T	Female	Mean age (sd)
HC	376	162	192	74.8 (5.8)
AD	251	140	114	75.3 (7.8)

2.2 Image Processing

Groupwise Registration. As our method defines features at the voxel level, it was necessary to transfer images into a common space. All native space images were rigidly and then affinely registered to a randomly chosen image, coalescing the registered images to update the template after each round of registrations. This was then followed by ten rounds of nonrigid registration to produce a final template in the groupwise space. All registrations were performed using the Niftyreg package [14].

Image Segmentation. All images were segmented into GM, white matter (WM), cerebrospinal fluid (CSF), and non-brain tissues components using the new segment module of SPM12 with the cleanup option set to maximum. A brain mask generated from the original structural image was then applied to the GM segmentations to further exclude any non-brain material.

Image Parcellation. The native space images were also anatomically parcellated into 83 regions. This was done with a novel label fusion algorithm [15] in a multi-atlas label propagation scheme. A library of 30 atlases manually labelled with 83 anatomical regions was used as a basis for the parcellation [16].

Atlas Construction. Unlike in other approaches using anatomical regions, features were defined at the level of the voxel rather than regions, requiring that all images share a common space. As kernels were constructed from the voxels within anatomical regions common across subjects, the parcellation defining the region was also required to be in the common space. However, our initial parcellations were in the native spaces of each subject. To combine these initial

parcellations in the groupwise space, the following procedure was used. First, all the parcellations were warped into the groupwise space, using the parameters from the native space of each image to the final groupwise template. Care was taken to preserve the integer labels in the parcellations during resampling. Finally to combine the individual parcellations, a consensus atlas was produced by majority voting among the set of N parcellations X to assign a single label l to each voxel v_i of the groupwise space $\mathbf{\Omega}$:

$$v_i, i \in \Omega = \arg\max_{l} \sum_{j=1}^{N} \begin{cases} 1, & \text{if } X_{i,j} = l \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

The pipeline to construct the atlas is summarised graphically in Fig. 1.

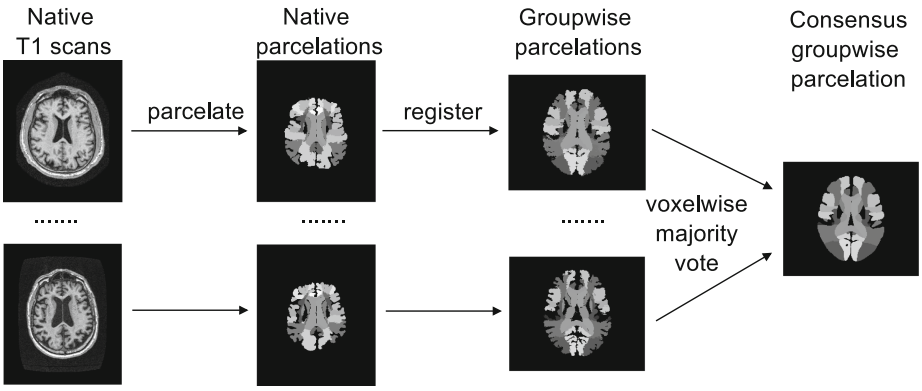


Fig. 1. Pipeline for constructing atlas in groupwise space

2.3 Gaussian Process Classification

Gaussian processes (GPs) provide a Bayesian, kernelised framework for solving both regression and classification problems. As an in depth explanation of GPs is beyond the scope of this paper, we refer the reader to [13] for a more theoretical treatment. Briefly, however, a GP (essentially a multivariate Gaussian) forms the prior on the value of a latent function f. For binary classification, the value of the latent function is linked to the probability of being in class $y, y \in \{-1, +1\}$ by a sigmoidal function. The GP is parameterised by a mean function $\mu(\mathbf{x})$ and a covariance kernel function $k(\mathbf{x}, \mathbf{x}')$ where $\mathbf{x}$ is a feature vector, whose elements represent voxel values in this case.

$$p(f(\mathbf{x}), f(\mathbf{x}') \sim \mathcal{N}(\mathbf{m}, \mathbf{K}), \text{ where } \mathbf{m} = \begin{bmatrix} \mu(\mathbf{x}) \\ \mu(\mathbf{x}') \end{bmatrix}, \mathbf{K} = \begin{bmatrix} k(\mathbf{x}, \mathbf{x}) & k(\mathbf{x}, \mathbf{x}') \\ k(\mathbf{x}', \mathbf{x}) & k(\mathbf{x}', \mathbf{x}') \end{bmatrix} \tag{2}$$

For classification, the non-Gaussian link function means that the posterior is also non-Gaussian so an approximation made be used. We make use of expectation propagation (EP) [17] as an approximation.

The GP prior is a function not only of the data but also of any hyperparameters θ that specify the form of the prior. We trained the GP by tuning the values of these hyperparameters to maximise the log likelihood of the training data, which can be done with standard gradient based optimisation algorithms. Once the hyperparameters have been set, predictions on unseen data are made by integrating across this optimised prior.

2.4 Gaussian Processes as Multimodal Kernel Methods

As Eq. 2 implies, GP classification belongs to the family of kernel methods. Hence a positive sum of valid kernels is a valid kernel, and a valid kernel multiplied by a positive scalar is also a valid kernel. The covariance between the ith and jth subject, $\mathbf{K}_{ij}$, is a kernel function k of the feature vectors for the ith and jth subject $\mathbf{x}_i$ and $\mathbf{x}_j$ and hyperparameters θ. For ARKs, the final kernel $\mathbf{K}$ is the weighted sum of 83 linear subkernels, each of which in turn is the dot product between the voxels within a particular anatomical region of the ith and jth image. These regions are defined using masks for each label derived from the groupwise atlas. This is illustrated in Fig. 2.

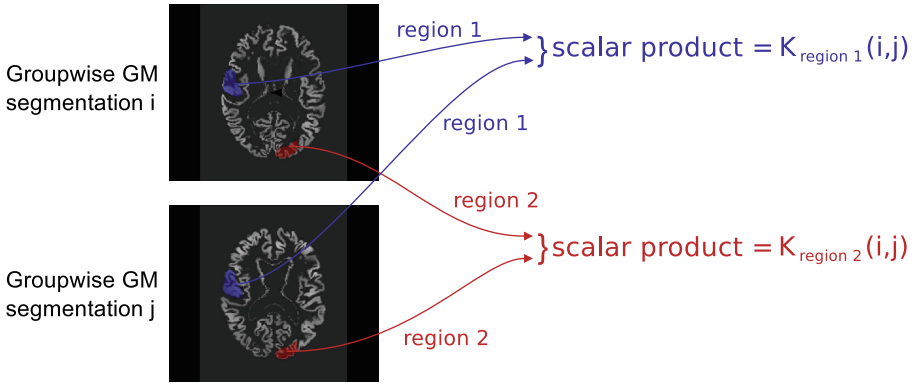


Fig. 2. Construction of anatomical regional kernels

The covariance hyperparameters are the weights of the subkernels α and bias term β, so the final kernel value $\mathbf{K}$ is given by

$$\mathbf{K}_{ij} = k(\mathbf{x}_i, \mathbf{x}_j) = \beta + \sum_{r=1}^{83} \alpha_r (\mathbf{x}_{i,r} \cdot \mathbf{x}_{j,r}) \tag{3}$$

where r indexes regions 1 to 83 and β is a bias term. There are thus 84 covariance hyperparameters: $\theta_{cov} = (\alpha_1, \alpha_2, \ldots, \alpha_{83}, \beta)$. All the above calculations are

carried out within the GPML toolkit[2], modified to take precomputed kernel matrices.

3 Results

We performed binary classification of all subjects as HC or AD. To generate results, we use a leave-one-out cross validation (LOOCV) across the entire set of 627 subjects. For the ARK formulation described above, the feature vectors $\mathbf{x}$ consist of voxel level data. For the purposes of comparison to existing methods, we also deploy two more conventional methods related to those introduced, representing opposite ends of the tradeoff between detail and use of prior anatomical information discussed in the introduction:

'voxels' method: This again uses voxel level data for the whole brain. However this is just used with a single kernel for the whole brain and no use of the atlas or anatomical prior information.

'regions' method: In place of voxel GM densities, this method takes the total GM volumes of each region as its features. These are normalised by the intra-cranial volume to control for variability in head size. The resulting feature vectors, of much lower dimensionality than either ARK or the voxels methods, are then used to build the single kernel.

3.1 Binary Accuracy

We compare the three methods by thresholding predicted probabilities at 0.5 and comparing to ground truth labels for HC or AD status. The resulting sensitivity, specificity and accuracy are shown in Table 2. We also show the area under the ROC curve (AUC), and a p-value for difference in accuracy with McNemar's test. The ARK formulation displays a greater accuracy and AUC then both competing methods. While the advantage over the voxels method is substantial, we do not quite have enough subjects and thus statistical power to show that it or the smaller advantage over regions is significant. We can, however visualise the effect of the ARK formulation across all the individual predictions.

Table 2. Binary accuracy summary

Method	Sens (%)	Spec (%)	Acc (%)	p vs ARK for acc	AUC
ARK	80.9	92.6	87.9	–	0.937
Voxels	73.7	93.9	85.8	0.166	0.914
Regions	80.1	91.0	86.6	0.409	0.9275

[2] http://www.gaussianprocess.org/gpml/code/matlab/doc/.

3.2 Individual Predictions

Figure 3 shows the *difference* in predicted $p(AD)$ between ARKs and each competing method for *all* subjects. Results are colour-coded so AD subjects are shown in red and HC ones in blue, and sorted by the value of the $p(AD)$ for the competing method. Hence blue (HC) subjects will be represented by a line extending left from the baseline, and red (AD) subjects by a line extending right, if ARKs improve the baseline classification. The plots also show how most subjects are correctly classified: The AD subjects mostly occupied the right hand side of the plots ($p(AD) > 0.5$) and the HC ones the left side of the plots.

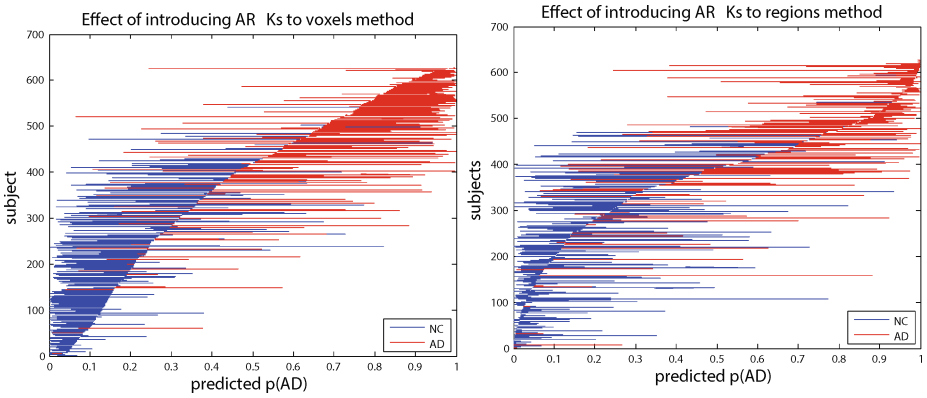


Fig. 3. Effects of ARKs on predictions for individual subjects

3.3 Interpretation of Hyperparameters

The optimised weights α tell us about the importance of the corresponding regions in the classification, and hence in AD. For each of the 627 sets of α, we normalise α so they represent a fraction of the total weight, than average each normalised weight across all folds. Only 14 regions have weights of more than 1 % of the total, shown in Fig. 3. These include temporal lobe regions frequently implicated in AD in studies such as [18], as well as the GM tissue adjacent to the temporal horn of the left lateral ventricle, which will be very sensitive to expansion of the horn. However, other structures much more widely distributed across the brain are also important in the classification, suggesting that atrophy may more quite widespread. The largest weight value is given to the right nucleus accumbens, and the left nucleus accumbens and right caudate are also given large weights, which may be a result of recently identified AD related atrophy in deeper structures [19] (Fig. 4).

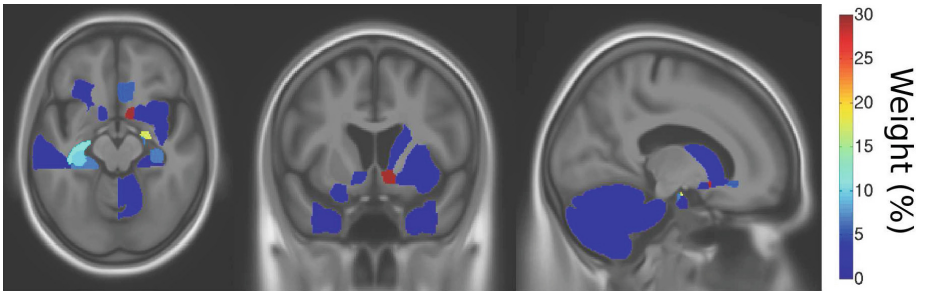


Fig. 4. Maps of regions with more than 1 % of total weight

4 Discussion

Our results show that ARKs successfully combine voxel level data with prior anatomical knowledge, offering a substantial accuracy improvement compared to voxel level data alone, and also offer a smaller improvement over features based on predefined regions. We are also able to show the improvements ARKs bring to individual subjects. Moreover, the kernel weights enhance model interpretability by showing new regions which may be involved in the AD process. The chief disadvantage of ARKs is speed of classifier training, due to the high dimensionality of the data and the large number of hyperparameters; however this is largely compensated for by the use of modified software that uses precomputed (sub)kernel matrices.

The method is quite general, and could also be applied to any type of training data where low level features and a parcellation are available in a common space, for example voxelwise cortical thickness data and a cortical atlas.

References

1. Klöppel, 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 Alzheimerss disease. Brain: J. Neurol. **131**(Pt 3), 681–689 (2008)
2. 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)
3. Fan, Y., Shen, D., Gur, R.C., Gur, R.E., Davatzikos, C.: COMPARE: classification of morphological patterns using adaptive regional elements. IEEE Trans. Med. Imaging **26**(1), 93–105 (2007)
4. Cuingnet, R., Gerardin, E., Tessieras, J., Auzias, G., Lehricy, 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 **56**(2), 766–781 (2011)
5. Young, J., Modat, M., Cardoso, M.J., Mendelson, A., Cash, D., Ourselin, S.: Accurate multimodal probabilistic prediction of conversion to Alzheimer's disease in patients with mild cognitive impairment. NeuroImage: Clin. **2**, 735–745 (2013)

6. 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)
7. Chu, C., Bandettini, P., Ashburner, J., Marquand, A., Kloeppel, S.: Classification of neurodegenerative diseases using Gaussian process classification with automatic feature determination. In: 2010 First Workshop on Brain Decoding: Pattern Recognition Challenges in Neuroimaging (WBD), pp. 17–20. IEEE (2010)
8. Liu, F., Zhou, L., Shen, C., Yin, J.: Multiple kernel learning in the primal for multi-modal Alzheimer's disease classification (2013). arXiv e-print 1310.0890
9. Gramfort, A., Thirion, B., Varoquaux, G.: Identifying predictive regions from fMRI with TV-L1 prior. In: Proceedings of the 2013 International Workshop on Pattern Recognition in Neuroimaging. PRNI 2013, pp. 17–20. IEEE Computer Society, Washington, DC (2013)
10. Cuingnet, R., Glaunès, J.A., Chupin, M., Benali, H., Colliot, O.: Spatial and anatomical regularization of SVM: a general framework for neuroimaging data. IEEE Trans. Pattern Anal. Mach. Intell. **35**(3), 682–696 (2013)
11. Sabuncu, M.R., Leemput, K.V.: The relevance voxel machine (RVoxM): a self-tuning Bayesian model for informative image-based prediction. IEEE Trans. Med. Imaging **31**(12), 2290–2306 (2012)
12. Neal, R.M.: Bayesian Learning for Neural Networks. Springer, New York (1996)
13. Rasmussen, C.E., Williams, C.K.I.: Gaussian Processes for Machine Learning. MIT Press, Cambridge (2006)
14. Modat, M., Ridgway, G.R., Taylor, Z.A., Lehmann, M., Barnes, J., Hawkes, D.J., Fox, N.C., Ourselin, S.: Fast free-form deformation using graphics processing units. Comput. Methods Programs Biomed. **98**(3), 278–284 (2010)
15. Cardoso, M., Modat, M., Ourselin, S., Keihaninejad, S., Cash, D.: Multi-STEPS: multi-label similarity and truth estimation for propagated segmentations. In: 2012 IEEE Workshop on Mathematical Methods in Biomedical Image Analysis (MMBIA), pp. 153–158 (2012)
16. Gousias, I.S., Rueckert, D., Heckemann, R.A., Dyet, L.E., Boardman, J.P., Edwards, A.D., Hammers, A.: Automatic segmentation of brain MRIs of 2-year-olds into 83 regions of interest. NeuroImage **40**(2), 672–684 (2008)
17. Minka, T.: Expectation propagation for approximate bayesian inference. In: Proceedings of the Seventeenth Conference Annual Conference on Uncertainty in Artificial Intelligence (UAI 2001), pp. 362–369. Morgan Kaufmann, San Francisco (2001)
18. Braak, H., Braak, E.: Staging of Alzheimer's disease-related neurofibrillary changes. Neurobiol. Aging **16**(3), 271–278 (1995)
19. Madsen, S., Ho, A., Hua, X., Saharan, P., Toga Jr, A., Jack, C., Weiner, M., Thompson, P.: 3D maps localize caudate nucleus atrophy in 400 Alzheimers disease, mild cognitive impairment, and healthy elderly subjects. Neurobiol. Aging **31**(8), 1312–1325 (2010)

Computer Aided Diagnosis

A Graph Based Classification Method for Multiple Sclerosis Clinical Forms Using Support Vector Machine

Claudio Stamile[1], Gabriel Kocevar[1], Salem Hannoun[1], Françoise Durand-Dubief[1,2], and Dominique Sappey-Marinier[1,3(✉)]

[1] CREATIS; CNRS UMR 5220; INSERM U1044, Université de Lyon, Université Lyon 1, INSA-Lyon, Villeurbanne, France
{claudio.stamile,gabriel.kocevar,salem.hannoun,francoise.durand-dubief, dominique.sappey-marinier}@creatis.insa-lyon.fr
[2] Service de Neurologie A, Hôpital Neurologique, Hospices Civils de Lyon, Bron, France
[3] CERMEP - Imagerie du Vivant, Université de Lyon, Bron, France

Abstract. With the development of advanced image acquisition and processing techniques providing better biomarkers for the characterization of brain diseases, automatic classification of biomedical imaging becomes an important field in research. Since brain neural network is one of the most complex network, graph theory constitutes a promising approach to characterize its connectivity properties. In this work, we applied this technique to diffusion tensor imaging data acquired in multiple sclerosis (MS) patients in order to classify their clinical forms. Support Vector Machine (SVM) algorithm in combination with graph kernel were used to classify 65 MS patients in three different clinical forms. Results showed high classification performances using both weighted and unweighted connectivity graphs, the later being more stable, and less dependent to the pathological conditions.

Keywords: Multiple sclerosis · SVM · Graph kernel · Clinical classification · Structural connectome

1 Introduction

Complex network analysis allows to describe highly structured data, simply through a geometric representation [1]. Such models have been used to study social behaviors [2], and have recently opened new perspectives in neuroscience, to study functional and structural brain connectivity using graph-derived metrics [3]. On one hand, analysis of the neural connections by functional magnetic resonance imaging (fMRI) provides networks, where nodes are active functional regions and links correspond to temporal functional correlations. This new approach gave the opportunity to characterize either cognitive impairments or pathological alterations caused by different brain diseases including Multiple Sclerosis (MS) [4,5]. On the other hand, brain structural connectivity based

K.K. Bhatia and H. Lombaert (Eds.): MLMMI 2015, LNCS 9487, pp. 57–64, 2015.
DOI: 10.1007/978-3-319-27929-9_6

on diffusion tensor imaging (DTI) data can be described using graph theory methods. Such structural networks are described by nodes, corresponding to segmented cortical regions, and links, reconstructed by tractography [6] from white matter (WM) fibers-tracts. Since structural connectome provides a fine description of anatomical connections between different cortical areas that could be modified by local [7] and/or diffuse [3] pathological mechanisms occurring in brain diseases such as MS. Thus, graph analysis provides a potential tool to better characterize MS disease and extract new brain biomarkers. Indeed, pathological events occurring in brain of MS patients constitute a rich source of open problems for image processing. This includes for instance, lesions segmentation algorithms [8], WM fiber-bundles analysis [9], new acquisition models [10] and others automatic algorithms.
Multiple sclerosis is the most frequent disabling neurological disease in young adults with a national prevalence of 95/100 000 in France [11]. MS is a demyelinating, inflammatory, chronic disease of the central nervous system. Disease onset is characterized by a first acute episode called clinically isolated syndrome (CIS), that evolves either into a relapsing-remitting (RR) course in about 85 % of cases or into a primary progressive (PP) course in the remaining 15 % of cases. RR patients will evolve into a secondary progressive (SP) course after several years. Today's neurologist challenge is to predict the individual patient evolution and response to therapy based on the clinical, biological and imaging markers available from disease onset. In this work, we will focus on the classification of MS patients in different groups of clinical forms. For the first time, we will try to solve this prognostic question using a computer-based method. Due to the unknown etiology of MS and the variability of the patients' clinical history, "model base" approaches are not suitable. This limitation could be overcome using "data-driven" approaches based on machine learning algorithms. Therefore, we propose a new fully automated method based on support vector machine (SVM) algorithm to classify MS clinical forms using patients structural connectivity information.

This paper is structured as follows. In Sect. 2, we provide a detailed description of our approach. In Sect. 3, we present the classification results. In Sect. 4 we discuss our results. Finally, in Sect. 5, we draw our conclusions.

2 Proposed Approach

2.1 Brain Structural Connectivity Graph

The connectivity graph of each subject was obtained by merging three different step. First, cortical and sub-cortical parcellation was obtained from T1-weighted MR images using FreeSurfer [12]. The obtained segmentation was used to label each voxel in one of the five tissue-type (cortical grey matter (GM), sub-cortical GM, WM, cerebrospinal fluid (CSF) and abnormal appearing tissue), and then, define the graph nodes. Second, pre-processing of diffusion images included correction of Eddy-current distortions [13] and skull stripping. Third, main diffusion

directions were estimated in each voxel using MRtrix spherical deconvolution algorithm [14]. Spherical harmonic (order of $L_{max} = 4$) was used to estimate both response function and orientation distribution function (ODF). Probabilistic streamline tractography algorithm [14] was applied to generate fiber-tracks in voxels labeled as WM voxels.

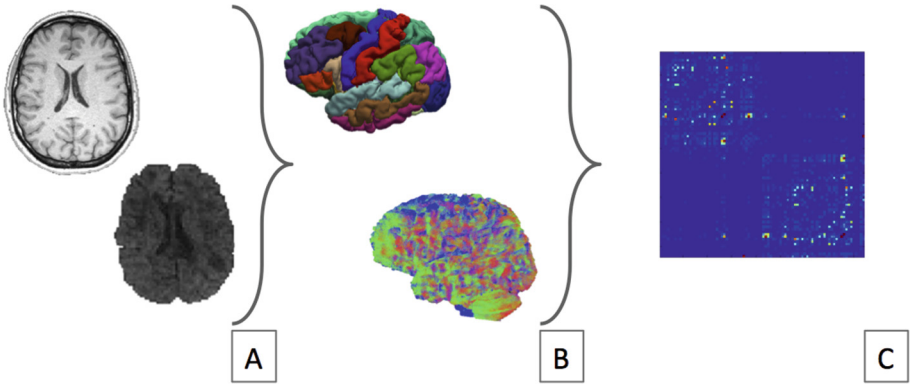


Fig. 1. Illustration showing the different steps of graph construction: (A) T1 and diffusion weighted MR images are used to generate cortical parcellation and fiber tractography (B), which are combined to generate connectivity matrix (C).

Symmetrical connectivity matrix $A \in \mathbb{N}_{+}^{q \times q}$ was then generated for each subject through the combination of GM segmentation and WM tractography. A schematic representation of graph construction steps is shown in Fig. 1. Let $a_{i,j}$ with $1 \leq i, j \leq q$ be an element of A, then $a_{i,j} = \Psi(i, j)$, where $\Psi : \mathbb{N}_1^2 \rightarrow \mathbb{N}$ denotes the number of fibers connecting the node i with the node j. The obtained connectivity matrix A is the representation of the weighted undirected graph $G = (V, E, \omega)$ where V ($|V| = q$) is the set containing the segmented GM brain regions, E is the graph edges set defined as:

$$E = \{\{i, j\} \mid \Psi(i, j) > 0 \; \forall \; 1 \leq i, j \leq q\}$$

and $\omega : E \rightarrow \Psi(E)$ is the weighted function that assigns at each edge $e \in E$ its weight. Roughly speaking this function is the same as Ψ but it is defined only on the element of the edges set E.

Starting from the weighted undirected graph $G = (V, E, \omega)$, we can generate an unweighted undirected graph $G^{'} = (V^{'}, E^{'})$ containing only the strongly connected regions respect to a given threshold $\gamma \in \mathbb{R}_{[0,1]}$. The graph binarization function $\Upsilon : G \rightarrow G^{'}$ performs the following mapping:

$$V^{'} = V$$

$$E^{'} = L(1, \ldots, T), \; T = \frac{(q^2 - q)\gamma}{2}$$

where L is the list of graph edges (E) sorted in ascending order of weight. In other words, the function Υ creates, starting from a weighted graph, an unweighted graph containing only the T strongest connections of G.

2.2 Classification Using SVM

Support Vector Machines (SVM) are a family of supervised classification algorithms [15]. The idea behind SVM classifier is to find the best hyperplane to separate data belonging to two different classes. More in detail, let $S = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a set of instances where $x_i \in \mathbb{R}^m$ and $y \in \{-1, 1\}^m$, a "soft margin" SVM classifier is based on the solution of the following optimization problem:

$$\begin{aligned} \underset{w,b}{\text{minimize}} \quad & \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{n}\epsilon_i \\ \text{subject to} \quad & y_i(w^T\phi(x_i) + b) \geq 1 - \epsilon_i,\ i = 1, \ldots, n \\ & \epsilon_i \geq 0,\ i = 1, \ldots, n \end{aligned}$$

Where ϵ is a relaxation variable of the optimization problem and C is the error penalization constant. The function $\phi(x_i)$ is mapping the feature vector x_i to an higher dimensional space.

The Lagrangian duality formulation of this problem is:

$$\begin{aligned} \underset{\alpha}{\text{maximize}} \quad & \sum_{i=1}^{n}\alpha_i - \frac{1}{2}\sum_{i,j=1}^{n} y_i y_j \alpha_i \alpha_j \langle\phi(x_i), \phi(x_j)\rangle \\ \text{subject to} \quad & 0 \leq \alpha_i \leq C,\ i = 1, \ldots, n \\ & \sum_{i=1}^{n}\alpha_i y_i = 0 \end{aligned}$$

where α_i are Lagrange multipliers. We can rewrite the inner product $\langle\phi(x), \phi(x)\rangle$ as a function $K(x, y) = (\phi(x)^T\phi(y))$ called kernel. Different kernel functions, mapping input space in higher dimensional space, are described in literature: polynomial, radial basic function (RBF), sigmoid function and others [16]. Since the weighted function ω of the generated graph G is such that $\forall e \in E\ \omega(e) > 0$, no negative loops are present in our graph. Thanks to this property, it is possible to apply the shortest-path graph kernel described in [17].

3 Classification Results

Sixty-five MS patients (24 males and 41 females, age = 39 ± 7 years (mean±sd)) including 24 RR, 24 SP, and 17 PP were included in this study. Local and national ethic committee (CPP Sud-Est IV and ANSM) approved this study and informed consent was collected before subject inclusion. Patients underwent

a MR examination on a 1.5T Siemens Sonata MR system with an 8-channel head-coil. The protocol consisted in a 3D T1-weighted sequence, acquired in the bi-commissural plane (isotropic $1\,mm^3$, TE/TR = 4/2000 ms), and a 2D-spin-echo DTI sequence (isotropic $2.5\,mm^3$, TE/TR= 86/6900 ms; 24 gradient-directions; $b = 1000\,s.mm^{-2}$). For each patient, the graph connectivity matrix was calculated using the method described in Sect. 2.1. Eighty-four anatomical regions were segmented and 500000 fibers were generated. Once all the connectivity matrices were computed and the graphs were generated, two different classification : first, using the weighted graphs classification (WGC) and second, the unweighted graphs classification (UGC). In case of UGC, the classification was performed using different threshold values γ. For each task, three different two-class classification ("RR vs PP", "RR vs SP", "PP vs SP") and one multi-class classifications ("RR vs PP vs SP") were performed using SVM with graph kernel. Generalization of classification performances was ensured by K-Fold cross validation [18] using leave-one-out criterion. The performances of both WGC and UGC were evaluated by calculating accuracy, precision and F-Measure [19] as reported in Table 1.

Table 1. Classification results using weighted (WGC) and unweighted (UGC, $\gamma = 0.75$) graphs.

	RR vs PP		RR vs SP		PP vs SP		RR vs PP vs SP	
	WGC	UGC	WGC	UGC	WGC	UGC	WGC	UGC
Accuracy (%)	82.9	68.3	64.6	66.7	51.2	70.7	35.4	47.7
Precision (%)	84.4	68.1	64.8	66.7	52.9	70.7	34.4	48.4
F-Measure (%)	80.3	68.1	64.6	66.7	51.5	70.7	33.7	47.5

The first classification task using the WGC reached the highest performance for "RR vs PP" classification with a F-Measure of 80.3 %. In contrast, the worst performance was obtained classifying "RR vs PP vs SP" with a F-Measure of 33.7 % (which is under the uncertainty level).
Results obtained by the second classification task using UGC for different γ values are illustrated in Fig. 2. The F-Measure values showed a large variability in the classification performances for $\gamma < 0.75$. In contrast, better performances and a greater stability were obtained for high threshold values ($0.75 \leq \gamma \leq 1$). In this range of γ values, UGC of "RR vs SP" and "PP vs SP" reached better performances than WGC, while WGC of "RR vs PP" remained better than UGC (Table 1). Similarly to the WGC, the multi-class UGC task ("RR vs PP vs SP") showed the worst performance with a F-Measure of 47.5 % ($\gamma = 0.75$). When comparing the average classification performances, no significant differences were found between WGC and UGC. While the performances of the two-class WGC are highly variable (sd(F-measure)= 12.6), especially for "RR vs PP" and "PP vs SP" classifications, UGC results showed a lower variability (sd(F-Measure)= 1.7).

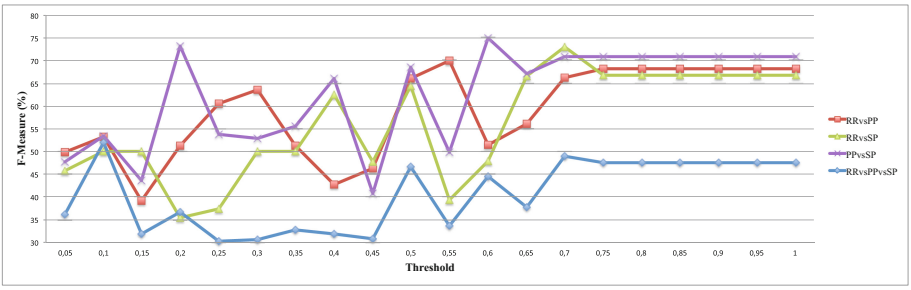


Fig. 2. Variations of F-Measure of unweighted graph classification with γ threshold values varying from 0 to 1.

4 Discussion

Similar average performances were obtained between the three clinical classifications using WGC and UGC. However, a large variability in performances was observed between the two-class clinical classifications when using WGC. In contrast, the performances of UGC for $\gamma > 0.75$ were lower than WGC but similar in between the two-class classifications.

From a methodological point of view, the two classification tasks, namely WGC and UGC, differ by the lost of knowledge concerning the number of fibers between edges in UGC due to binarization. This methodological difference may explain the better performances obtained with WGC compared to UGC when classifying "RR vs PP" and suggest that the fibers number plays a significant role in the classification. RR and PP are both starting clinical forms of MS, that are characterized by very different pathological processes. RR patients are subject to remitting focal lesions mainly due to inflammatory processes while PP patients are subject to diffuse and progressive degenerative mechanism [20]. These neurodegenerative processes lead to severe axonal loss that is concordant with the poor clinical status of PP patients. Thus, the greater fiber loss occurring in PP patients compared to RR patients may lead to a significant difference in fibers number distribution which can help the WGC to discriminate RR from PP clinical forms. Nevertheless, this finding may also reflect a limitation of tractography algorithms to reconstruct short associative fibers, particularly in presence of diffuse WM tissue damages occurring in PP forms.

In summary the WGC provides, in average, equal performance compared to UGC. However, WGC is associated with a large variability of its performances between the three clinical classifications. Fortunately these limits, could be overcome by using the UGC method. Indeed, despite the information reduction due to thresholding, the SVM algorithm in combination with graph kernel allows to classify MS patients with acceptable performances.

5 Conclusion

In this paper, we proposed a graph-based method to classify MS patients according to their clinical forms. Graph theory was applied to describe brain network topology and SVM classification was performed using weighted and unweighted graphs. The main result of this study showed that the use of "data-driven" methods like machine learning algorithm are suitable in environments where it is difficult to build a descriptive model like in MS disease. Moreover, the high performances obtained when classifying "RR vs PP", the two starting MS clinical forms, make our method a potential tool to allow a better prediction of disability progression in MS patients.

In conclusion, SVM classifiers based on sensitive and global image biomarkers, such as structural graphs based on DTI data, may constitute a new and sensitive tool for brain disease classification.

Acknowledgements. Claudio Stamile is funded by an EU-funded FP7-PEOPLE-2012-ITN project 316679 TRANSACT. This work is supported by the French National Research Agency (ANR) within the national program "Investissements d'Avenir" through the OFSEP project (ANR-10-COHO-002).

References

1. Shuman, D.I., Narang, S.K., Frossard, P., Ortega, A., Vandergheynst, P.: The emerging field of signal processing on graphs. IEEE Sign. Proces. Mag. **5**, 83–98 (2013)
2. Klovdahl, A.S., Graviss, E.A., Yaganehdoost, A., Ross, M.W., Wanger, A., Adams, G.J., Musser, J.M.: Networks and tuberculosis: an undetected community outbreak involving public places. Soc. Sci. Med. **52**(5), 681–694 (2001)
3. Achard, S., Delon-Martin, C., Vrtes, P.E., Renard, F., Schenck, M., Schneider, F., Heinrich, C., Kremer, S., Bullmore, E.T.: Hubs of brain functional networks are radically reorganized in comatose patients. Proc. Natl. Acad. Sci. U.S.A. **109**(50), 20608–20613 (2012)
4. Filippi, M., Van den Heuvel, M.P., Fornito, A., He, Y., Hulshoff Pol, H.E., Agosta, F., Comi, G., Rocca, M.A.: Assessment of system dysfunction in the brain through MRI-based connectomics. Lancet Neurol. **12**, 1189–1199 (2013)
5. Bassett, D.S., Bullmore, E., Verchinski, B.A., Mattay, V.S., Weinberger, D.R., Meyer-Lindenberg, A.: Hierarchical organization of human cortical networks in health and schizophrenia. J. Neurosci. **28**, 9239–9248 (2008)
6. Mori, S., Crain, B.J., Chacko, V.P., van Zijl, P.C.: Three-dimensional tracking of axonal projections in the brain by magnetic resonance imaging. Ann. Neurol. **45**(2), 265–269 (1999)
7. Crofts, J.J., Higham, D.J., Bosnell, R., Jbabdi, S., Matthews, P.M., Behrens,T.E.J., Johansen-Berg, H.: Network analysis detects changes in the contralesional hemisphere following stroke. Neuroimage **54**, 161–169 (2011)
8. Lladó, X., Oliver, A., Cabezas, M., Freixenet, J., Vilanova, J.C., Quiles, A., Valls, L., Ramió-Torrentà, L., Rovira, A.: Segmentation of multiple sclerosis lesions in brain MRI: a review of automated approaches. Inf. Sci. **186**, 164–185 (2012)

9. Stamile, C., Kocevar, G., Cotton, F., Durand-Dubief, F., Hannoun, S., Frindel, C., Rousseau, D., Sappey-Marinier, D.: Detection of longitudinal DTI changes in multiple sclerosis patients based on sensitive WM fiber modeling. In: ISMRM, Toronto, Canada, May-Jun 2015
10. Zhang, H., Schneider, T., Wheeler-Kingshott, C.A., Alexander, D.C.: NODDI: practical in vivo neurite orientation dispersion and density imaging of the human brain. Neuroimage **61**(4), 1000–1016 (2012)
11. Fromont, A., Binquet, C., Sauleau, E.A., Fournel, I., Bellisario, A., Adnet, J., Weill, A., Vukusic, S., Confavreux, C., Debouverie, M.: Geographic variations of multiple sclerosis in France. Brain **133**(7), 1889–1899 (2010)
12. Fischl, B., Van der Kouwe, A., Destrieux, C., Halgren, E., Segonne, F., Salat, D.H., Busa, E., Seidman, L.J., Goldstein, J., Kennedy, D., Caviness, V., Makris, N., Rosen, B., Dale, A.M.: Automatically parcellating the human cerebral cortex. Cereb. Cortex **14**, 11–22 (2004)
13. Jenkinson, M., Beckmann, C.F., Behrens, T.E.J., Woolrich, M.W., Smith, S.M.: FSL. Neuroimage **62**(2), 782–790 (2012)
14. Tournier, J., Calamante, F., Connelly, A.: MRtrix: diffusion tractography in crossing fiber regions. IJIST **22**(1), 53–66 (2012)
15. Cortes, C., Vapnik, V.: Support-vector networks. Mach. Learn. **20**(3), 273–297 (1995)
16. Gärtner, T.: A survey of kernels for structured data. ACM SIGKDD Explor. Newsl. **5**(1), 49–58 (2003)
17. Borgwardt, K.M., Kriegel, H.P.: Shortest-path kernels on graphs. In: Proceedings of the International Conference on Data Mining, pp. 74–81 (2005)
18. Kohavi, R.: A study of cross-validation and bootstrap for accuracy estimation and model selection. IJCAI **14**(2), 1137–1145 (1995)
19. Powers, D.: Evaluation: from precision, recall and f-factor to roc, informedness, markedness & correlation (Technical report), Adelaide, Australia (2007)
20. Zivadinov, R., Cox, J.L.: Neuroimaging in multiple sclerosis. Int. Rev. Neurobiol. **79**, 449–474 (2007)

Classification of Alzheimer's Disease Using Discriminant Manifolds of Hippocampus Shapes

Mahsa Shakeri[1,2], Hervé Lombaert[3], and Samuel Kadoury[1,2(✉)]

[1] MEDICAL, Polytechnique Montréal, Montréal, Québec, Canada
mahsa.shakeri@polymtl.ca, samuel.kadoury@polymtl.ca
[2] CHU Sainte-Justine Hospital Research Center, Montréal, Québec, Canada
[3] Asclepios Team, Inria Sophia-Antipolis Méditerranée, Sophia-antipolis, France
herve.lombaert@inria.fr

Abstract. Neurodegenerative pathologies, such as Alzheimer's disease, are linked with morphological alterations of subcortical structures which can be assessed from medical imaging and biological data. Recent advances in machine learning have helped to improve classification and prognosis rates. We present here a classification framework for Alzheimer's disease which extracts triangulated surface meshes from segmented binary maps in MRI, and establishes reliable point-to-point correspondences among a population of hippocampus 3D surfaces using their spectral representation. Morphological changes between groups are detected using a manifold learning algorithm based on Grassmannian kernels in order to assess similarity between shape topology in control normals and patients. A second manifold using discriminant embeddings is then generated to maximize the class separability between three clinical groups recognized in dementia. We test the method to classify 47 subjects with Alzheimers Disease (AD), 47 with mild cognitive impairment (MCI) and 47 healthy controls enrolled in a clinical study. Classification rates compare favorably to standard classification methods based on SVM and traditional manifold learning methods evaluated on the same database.

1 Introduction

Alzheimer's disease (AD) is the most common form of dementia, with an incidence that doubles every five years after the age of 65 [2]. As life expectancy increases, the number of AD patients increases accordingly, which causes a heavy socioeconomic burden. It is expected that treatment decisions will greatly benefit from diagnostic and prognostic tools that identify individuals likely to progress to dementia sooner. This is especially important in individuals with mild cognitive impairment (MCI), who present a conversion rate of approximately 15 % per year. Towards this end, neuroimaging datasets for AD including magnetic resonance imaging (MRI) and other types of biomarkers have shown considerable promise to detect longitudinal changes in subjects scanned repeatedly over time [13], by offering rich information on the patients morphometric and anatomical profiles. Their use stems from the premise that longitudinal changes may be

K.K. Bhatia and H. Lombaert (Eds.): MLMMI 2015, LNCS 9487, pp. 65–73, 2015.
DOI: 10.1007/978-3-319-27929-9_7

more reproducible and more precisely measured with MRI and other parameters such as in clinical scores, cerebrospinal fluid (CSF), or proteomic assessments.

A number of studies reported structural changes in the hippocampus, parahippocampal gyrus, cingulate, and other brain regions in both MCI and AD patients [12]. Other studies have used intensity information to discriminate elderly normal controls (NC) with patients inflicted with AD or mild cognitive impairment (MCI), based on T1-weighted MRI [6]. Previous machine learning algorithms using MRI were based on traditional morphometric measures, such as subcortical volume or shape descriptors of brain structures [3] and their change over time [5]. These were based on finding a low-dimensional representation of complex and high-dimensional data using principal component analysis (PCA) and multidimensional scaling (MDS). However these methods are typically linear, making it easy to transform data from image space into the learned subspace, but lacks the ability to process irregular or abnormal structures, which tend to follow non-linear patterns of variation. To cope with this limitation, manifold learning methods on the other hand tend to better model highly non-linear data, such as from neuroimaging datasets [1]. Recently, discriminant embeddings exploit within and between-class similarities to establish correspondences between disparate data, thereby offering a more accurate relationship of subtile structural alterations in AD.

The objective of this study is to propose a classifier which distinguishes NC subjects from patients with MCI and patients afflicted with AD. First, segmented hippocampus shapes from MRI are matched between each other using a spectral representation of the 3D mesh surface of the sub-cortical surface in order to have one-to-one vertex correspondences between hippocampus shapes throughout a population. Once a training set of hippocampus shapes is created for three clinical relevant groups (NC, MCI, AD), a discriminant manifold based on Grassmannian kernels is trained to maximize the separation between these three groups and improve the classification accuracy for any unseen MRI, which can be processed by mapping the segmented hippocampus onto the trained manifold. The main contribution of this paper is to develop a hippocamus classification approach based on their spectral representation which is classified in the Grassmannian space.

2 Methods

2.1 Hippocampus Shape Alignment

In the first step, segmented binary masks obtained from diagnostic T1-weighted MRI are processed to the same image orientation and isotropic voxel sizes, and then converted into 3D triangulated surfaces using the marching cube algorithm. A Gaussian smoothing process is subsequently applied on each surface in order to remove surface irregularities. Then, a reference surface is defined in an iterative process, and all triangulated surfaces are aligned to this reference using a rigid registration algorithm. In order to establish the point-to-point correspondences

across all surfaces, each mesh is matched to a randomly selected reference surface using a spectral matching algorithm as proposed in [8].

The matching between two surfaces S_i and S_j of the hippocampus from two separate subjects is conducted in a two-step process. In the first step, an initial transformation is calculated between the two surfaces, followed by a second step to establish a smooth map between the two meshes based on a diffeomorphic mapping [7]. First, the spectrums of the meshes S_i and S_j are computed according to spectral representation theory. Meshes are described by their principal eigenmodes following an eigendecomposition of their respective Laplacian matrix L. In order to add robustness to the feature matching process, the mean curvature at each point of the mesh defined as $C(i) = 0.5 * (C_{min} + C_{max})$ are calculated, where the principal curvatures C_{min} and C_{max} are estimated as the minimum and maximum curving degrees of a mesh S, respectively. Hence, the mean curvature of C is computed as $\{C(1), C(2), \cdots, C(n)\}$, where n is the number vertices. We incorporate these features in the weighting of the nodes of the spectral graph $\mathcal{G}$ by computing the exponential of the mean curvature, and defining the graph Laplacian as $\tilde{L} = \mathcal{G}L$, where

$$\mathcal{G} = P^{-1}(\exp(diag(\{C(1), C(2), \cdots, C(n)\})))^{-1} \tag{1}$$

and P is the diagonal node degree matrix integrating distance weights. Once meshes are described in the spectral domain, the first e eigenvectors associated with non-zero eigenvalues are chosen to define the spectral representations $\tilde{S}_i$ and $\tilde{S}_j$. After reordering and sign adjustment [7] of the resulting spectrums $\tilde{S}_i$ and $\tilde{S}_j$, we perform non-rigid alignment of the spectral coordinates using Coherent Point Drift (CPD) [9]. The CPD approach finds a continuous transformation between the surfaces $\tilde{S}_i$ and $\tilde{S}_j$ in the spectral domain. Once the two spectral representations are aligned, the point-by-point correspondences between two meshes could be directly established in the Euclidean space, such that the two closest points in the spectral domain are considered as corresponding points in the Euclidean space. Thus, the correspondence map c between S_i and S_j is established with a simple nearest-neighbor search in spectral domain.

It was shown in [8] that incorporating extra features might create discontinuities in the correspondence map c. As a solution, a diffeomorphic matching is applied to find the final map between two shapes. This is obtained by defining an association graph composed of the set of vertices and edges, based on the initial set of correspondence links. The graph Laplacian operator is applied on the resulting graph, followed by a spectral decomposition to produce a shared set of eigenvectors, from which the first and last eigenvalues are used to obtain one-to-one vertex correspondences between the mesh vertices. This procedure is repeated for all training meshes in the three groups of the database, with (1) normal controls, (2) MCI patients and (3) AD patients.

2.2 Learning the Discriminant Grassmannian Manifold

Manifold learning algorithms are based on the premise that data are often of artificially high dimension and can be embedded in a lower dimensional space.

However the presence of outliers and multi-class information can on the other hand affect the discrimination and/or generalization ability of the manifold. We propose to learn the optimal separation between three classes (1) normal controls, (2) MCI patients and (3) AD patients, by using a discriminant graph-embedding based on Grassmannian manifolds for the classification problem initially proposed in [4]. Each sample mesh surface S, which vertices has been rearranged using the alignment method in 2.1, can be viewed as the set of low-dimensional m subspaces of $\mathbb{R}^n$ on a Grassmannian manifold and represented by orthonormal matrices, each with a size of $n \times m$, with n the higher dimensionality of vertices defined earlier. Two points on a Grassmannian manifold are equivalent if one can be mapped into the other one by a $m \times m$ orthogonal matrix. In this work, similarity between two surfaces (S_i, S_j) on the manifold is measured as a combination of projection and canonical correlation Grassmannian kernels $\mathbb{K}_{i,j}$ defined in the Hilbert Space. By describing different features of the hippocampus shape with each kernel, $\mathbb{K}_{i,j}$ can improve discriminatory accuracy between shapes.

In order to effectively discover the low-dimensional embedding, it is necessary to maintain the local structure of the data in the new embedding. The structure $G = (\boldsymbol{V}, \boldsymbol{W})$ is an undirected similarity graph, with a collection of nodes $\boldsymbol{V}$ connected by edges, and the symmetric matrix $\boldsymbol{W}$ with elements describing the relationships between the nodes. The diagonal matrix $\boldsymbol{D}$ and the Laplacian matrix $\boldsymbol{L}$ are defined as $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$, with $\boldsymbol{D}(i,i) = \sum_{j \neq i} \boldsymbol{W}_{ij} \forall i$. Here, N labelled points $\mathbb{S} = \{(S_i, c_i)\}_{i=1}^{N}$ are generated from the underlying manifold $\mathcal{M}$, where c_i denotes the label (NC, MCI or AD). The task at hand is to maximize a measure of discriminatory power by mapping the underlying data into a vector space, while preserving similarities between data points in the high-dimensional space. Discriminant graph-embedding based on locally linear embedding (LLE) [11] uses graph-preserving criterions to maintain these similarities, which are included in a sparse and symmetric $N \times N$ matrix, denoted as M.

Within and Between Similarity Graphs: In our work, the geometrical structure of $\mathcal{M}$ can be modeled by building a within-class similarity graph $\boldsymbol{W}_w$ for hippocampus of the same group and a between-class similarity graph $\boldsymbol{W}_b$, to separate hippocampus from the three classes. When constructing the discriminant LLE graph, elements are partitioned into $\boldsymbol{W}_w$ and $\boldsymbol{W}_b$ classes. The intrinsic graph G is first created by assigning edges only to samples of the same class (ex: MCI). The local reconstruction coefficient matrix $M(i,j)$ is obtained by minimizing:

$$\min_{M} \sum_{j \in \mathcal{N}_w(i)} \|S_i - M(i,j) S_j\|^2 \quad \sum_{j \in \mathcal{N}_w(i)} M(i,j) = 1 \quad \forall i \tag{2}$$

with $\mathcal{N}_w(i)$ as the neighborhood of size k_1, within the same region as point i (e.g. hippocampus from MCI patient). Each sample is therefore reconstructed only from 3D meshes of the same clinical group. The local reconstruction coefficients are incorporated in the within-class similarity graph, such that the matrix $\boldsymbol{W}_w$ is defined as:

$$W_w(i,j) = \begin{cases} (M + M^T - M^T M)_{ij}, & \text{if } S_i \in \mathcal{N}_w(S_j) \text{ or } x_j \in \mathcal{N}_w(S_i) \\ 0, & \text{otherwise.} \end{cases} \quad (3)$$

Conversely, the between-class similarity matrix $\boldsymbol{W}_b$ depicts the statistical properties to be avoided in the optimization process and used as a high-order constraint. Distances between healthy and pathological samples are computed as:

$$W_b(i,j) = \begin{cases} 1/k_2, & \text{if } S_i \in \mathcal{N}_b(S_j) \text{ or } S_j \in \mathcal{N}_b(S_i) \\ 0, & \text{otherwise} \end{cases} \quad (4)$$

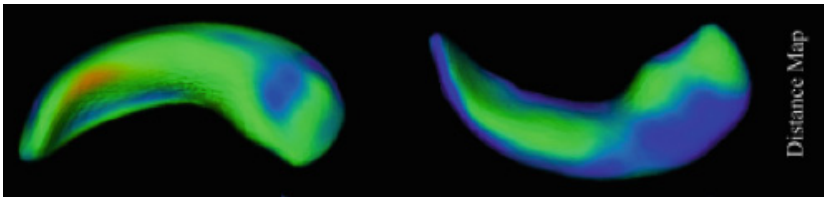
with $\mathcal{N}_b$ containing k_2 neighbors having different class labels from the ith sample. The objective is to transform points to a new manifold $\mathcal{M}$ of dimensionality d, i.e. $S_i \rightarrow y_i$, by mapping connected samples from the same group in $\boldsymbol{W}_w$ as close as possible to the class cluster, while moving NC, MCI and AD meshes of $\boldsymbol{W}_b$ as far away from one another. This results in optimizing the objective functions:

$$f_1 = \min \frac{1}{2} \sum_{i,j} (y_i - y_j)^2 W_w(i,j) \qquad f_2 = \max \frac{1}{2} \sum_{i,j} (y_i - y_j)^2 W_b(i,j) \quad (5)$$

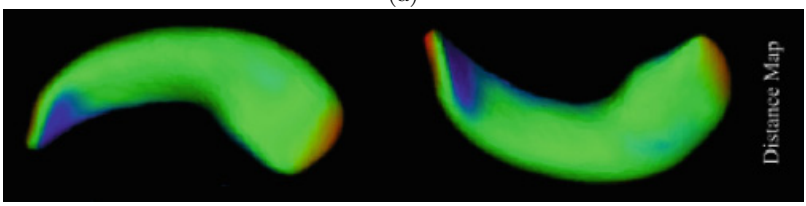
Supervised Manifold Learning: The optimal projection matrix, mapping new points to the manifold, is obtained by simultaneously maximizing class separability and preserving interclass manifold property, as described by the objective functions in Eq.(5). Assuming points on the manifold are known as similarity measures given by the Grassmannian kernel $\mathbb{K}_{i,j}$, a linear solution can be defined, i.e., $y_i = (\langle \alpha_1, S_i \rangle, \ldots, \langle \alpha_r, S_i \rangle)^T$ for the r largest eigenvectors with $\alpha_i = \sum_{j=1}^{N} a_{ij} S_j$. Defining the coefficient $\boldsymbol{A}_l = (a_{l1}, \ldots, a_{lN})^T$ and kernel $\boldsymbol{K}_i = (k_{i1}, \ldots, k_{iN})^T$ vectors, the output can be described as $y_i = \langle \alpha_l, S_i \rangle = \boldsymbol{A}_l^T \boldsymbol{K}_i$. By replacing the linear solution in the minimization and maximization of the between- and within-class graphs, the optimal projection matrix $\mathbb{A}$ is acquired from the optimization of the function as proposed in [4]. The proposed algorithm uses the points on the Grassmannian manifold implicitly (i.e., via measuring similarities through a kernel) to obtain a mapping $\mathbb{A}$. The matrix maximizes a quotient similar to discriminant analysis, while retaining the overall geometrical structure. Hence for any new segmented surface mesh S_q, a manifold representation can be obtained using the kernel function based on S_q and mapping $\mathbb{A}$.

3 Experiments and Results

We used the Alzheimer's Disease Neuroimaging Initiative (ADNI) database with 1.5 or 3.0 T structural MR images (adni.loni.usc.edu). For this study, a subset of baseline 1.5 T MR images is used including 47 normal controls (NC), 47 AD patients, and 47 individuals with MCI. The three groups are matched approximately by age and gender (NC with a mean age of 76.7 ± 5.4, 23 male; AD with



(a)



(b)

Fig. 1. (a) Distance maps of left and right hippocampal shape deformations in AD patients compared with normal controls. (b) Distance maps of left and right hippocampal shape deformations in MCI patients compared with normal controls.

a mean age of 77.4 ± 7.2, 21 males; MCI with a mean age of 75.0 ± 6.9, 28 males). Additional post-processing steps were performed on the MR images to correct certain image artifacts and to enhance standardization across sites and platforms. The post-processing steps include gradient non-linearity correction, intensity inhomogeneity correction, bias field correction, and phantom-based geometrical scaling to remove calibration errors. Here, we use these processed images. Left and right hippocampi were segmented using FSL-FIRST automatic segmentation [10] and visual inspection was performed on the output binary masks to ensure the quality of the segmentation. Figure 1 shows the shape differences in the left and right hippocampus between NC, MCI and AD.

The optimal size was found at $k_1 = 7$ for within-class neighborhoods ($\mathcal{N}_w$), and $k_2 = 4$ for between-class neighborhoods ($\mathcal{N}_b$). The optimal manifold dimensionality was set at $d = 5$, when the trend of the nonlinear residual reconstruction error curve stabilized for the entire training set. Figure 2 shows the resulting manifold with embedded hippocampus shapes which can be clearly identified into three separate groups, due to the discriminative nature of the framework. Table 1 presents accuracy, sensitivity and specificity results for SVM (nonlinear RBF kernel), LLE and the proposed method between three clinically relevant pairs of diagnostic groups (NC/AD, NC/MCI, MCI/AD). The classifier performance was obtained by repeating 100 times a random selection of samples, using 75 % of the data for training and 25 % for testing in each run. Results show a significant improvement using the discriminant manifold embedding compared to standard approaches. It also illustrates that increased accuracy can be achieved using the discriminant embedding with combined kernel ($\alpha_1 = 1, \alpha_2 = 5$), which suggests the benefit of extracting complementary features from the dataset for classification purposes compared to different types of classification models (SVM, LLE).

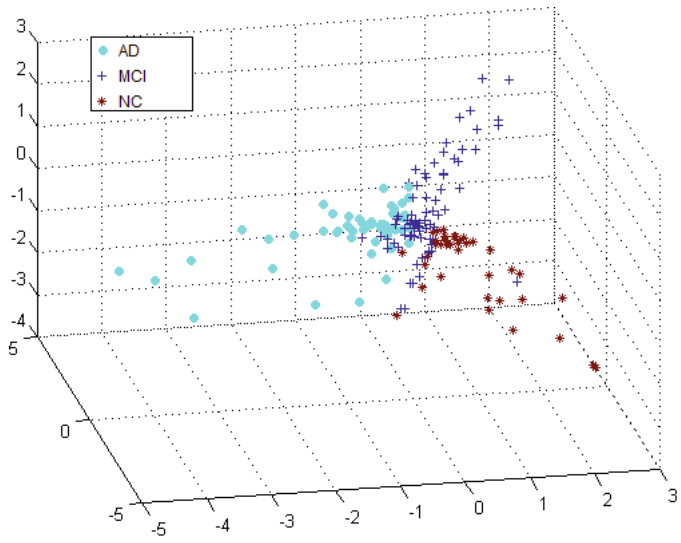


Fig. 2. Resulting manifold embedding with low-dimensional coordinates of samples points taken from the NC, MCI and AD groups.

Table 1. Classification results for the classification of NC, MCI and AD patients from segmented hippocampal regions. We compare a standard SVM classification approach, with a single LLE method and the proposed discriminant LLE method.

	NC/AD			NC/MCI			MCI/AD			All groups		
	SVM	LLE	DLLE	SVM	LLE	DLLE	SVM	LLE	DLLE	SVM	LLE	DLLE
Sensitivity tp/(tp+fn)	0.75	0.84	0.90	0.58	0.61	0.69	0.50	0.57	0.60	0.61	0.67	0.73
Specificity tn/(tn+fp)	0.69	0.77	0.85	0.62	0.70	0.77	0.57	0.61	0.67	0.62	0.69	0.77
Overall accuracy	0.72	0.79	0.88	0.60	0.65	0.72	0.54	0.58	0.65	0.62	0.67	0.74

4 Conclusion

Our main contribution consists in describing morphometric variations of the hippocampus in a discriminant nonlinear graph embedding with Grassmannian manifolds to detect the presence of Alzheimer's disease. A spectral matching process based on the eigendecomposition of the Laplacian matrix of hippocampus shapes extracted from a dataset of MRI images enabled to establish one-to-one correspondences in mesh vertices. This is critical to construct a reliable training set of sub-cortical shapes from various pathological groups and normal controls. A manifold embedding including intrinsic and penalty graphs measuring similarity within clinical relevant groups and between NC, MCI and AD patients, respectively, was trained to differentiate between the different hippocampus shapes. A combination of canonical correlation kernels creates a secondary manifold to simplify the deviation estimation from normality, improving detection of pathology compared to standard LLE. Experiments show the need of nonlinear embedding of the learning data, and the relevance of the proposed

method for stratifying different stages of dementia progression. In the context of Alzheimer's disease, the method can improve for the early detection of the disease with promising classification rates based on ground-truth knowledge. Future work will compare results to volumetric measurements and improve the deviation metric using high-order tensorization and investigate into fully automated hippocampus segmentation, as it can affect the precision of the spectral correspondence process.

References

1. Aljabar, P., Wolz, R., Srinivasan, L., Counsell, S.J., Rutherford, M.A., Edwards, A.D., Hajnal, J.V., Rueckert, D.: A combined manifold learning analysis of shape and appearance to characterize neonatal brain development. IEEE Trans. Med. Imaging **30**(12), 2072–2086 (2011)
2. Bain, L., Jedrziewski, K., Morrison-Bogorad, M., et al.: Healthy brain aging: a meeting report from the Sylvan M. cohen annual retreat of the university of pennsylvania institute on aging. alzheimers. Dementia **4**, 443–446 (2008)
3. Chupin, M., Hammers, A., Liu, R.S., 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**(3), 749–761 (2009)
4. Harandi, M., Sanderson, C., et al.: Graph embedding discriminant analysis on grassmannian manifolds for improved image set matching. In: CVPR. p. 2705 (2011)
5. Leow, A.D., Yanovsky, I., Chiang, M.C., Lee, A.D., Klunder, A.D., Lu, A., Becker, J.T., Davis, S.W., Toga, A.W., Thompson, P.M.: Statistical properties of jacobian maps and the realization of unbiased large-deformation nonlinear image registration. IEEE Trans. Med. Imaging **26**(6), 822–832 (2007)
6. Li, Y., Wang, Y., Wu, G., Shi, F., Zhou, L., Lin, W., Shen, D., Initiative, A.D.N., et al.: Discriminant analysis of longitudinal cortical thickness changes in alzheimer's disease using dynamic and network features. Neurobiology of aging **33**(2), 427–e15 (2012)
7. Lombaert, H., Grady, L., Polimeni, J.R., Cheriet, F.: FOCUSR: feature oriented correspondence using spectral regularization-a method for precise surface matching. IEEE Trans. Pattern Anal. Mach. Intell **35**(9), 2143–2160 (2013)
8. Lombaert, H., Sporring, J., Siddiqi, K.: Diffeomorphic spectral matching of cortical surfaces. In: Gee, J.C., Joshi, S., Pohl, K.M., Wells, W.M., Zöllei, L. (eds.) IPMI 2013. LNCS, vol. 7917, pp. 376–389. Springer, Heidelberg (2013)
9. Myronenko, A., Song, X., Miguel, A.C.: Non-rigid point set registration: coherent point drift. IEEE Trans. Pattern Anal. Mach. Intell **32**(12), 2262–2275 (2009)
10. Patenaude, B., Smith, S.M., Kennedy, D.N., Jenkinson, M.: A bayesian model of shape and appearance for subcortical brain segmentation. Neuroimage **56**(3), 907–922 (2011)
11. Roweis, S., Saul, L.: Nonlinear dimensionality reduction by locally linear embedding. Science **290**, 2323–2326 (2000)

12. Visser, P., Verhey, F., Hofman, P., Scheltens, P., Jolles, J.: Medial temporal lobe atrophy predicts alzheimer's disease in patients with minor cognitive impairment. J. Neurol. Neurosurg. Psychiatry **72**(4), 491–497 (2002)
13. Wyman, B.T., Harvey, D.J., Crawford, K., Bernstein, M.A., Carmichael, O., Cole, P.E., Crane, P.K., DeCarli, C., Fox, N.C., Gunter, J.L., et al.: Standardization of analysis sets for reporting results from ADNI MRI data. Alzheimer's Dementia **9**(3), 332–337 (2013)

Transfer Learning for Prostate Cancer Mapping Based on Multicentric MR Imaging Databases

Rahaf Aljundi[1(✉)], Jérôme Lehaire[1,2], Fabrice Prost-Boucle[1,2], Olivier Rouvière[2], and Carole Lartizien[1]

[1] Université de Lyon, CREATIS; CNRS UMR5220; INSERM U1044; INSA-Lyon; Université Lyon 1, Villeurbanne, France
Rahaf.Aljundi@gmail.com, Carole.Lartizien@creatis.insa-lyon.fr
[2] INSERM, U1032, LabTau; Université de Lyon, Lyon, France

Abstract. This paper addresses the issue of fusing datasets coming from different imaging protocols or scanners to boost the performance of computer aided diagnostic system. We present novel contributions in the field of subspace alignment methods that are part of domain adaptation framework. We first introduce a simple approach based on scaling the features of the different distribution and accounting for the class information. We also extend an unsupervised landmark based approach that has been recently developed to the supervised setting. These methods are evaluated in the context of prostate cancer screening based on two patient MRI databases acquired on different scanners. We demonstrate promising performance of the scaling based method when both databases contain similar number of annotated samples, and stable performance of the landmark based method even with unbalanced datasets.

Keywords: Transfer learning · Computer-aided detection system

1 Introduction

Computer-aided diagnosis (CAD) has become a major research subject in different domains of cancer imaging to assist radiologists during their diagnostic task by providing information on the location and characterization (malignancy score) of suspicious regions of interest. These algorithms learn a multiclass (mostly binary) decision model in a multidimensional feature space based on training samples from the different classes of interest. Diagnostic performance of such decision support systems is highly impacted by the quality of the training database that should contain a large number of correctly annotated cases of all classes. Such a condition is not easily met in clinical practice. First, images might not be acquired from the same imaging protocols or the same scanner model. This point is especially true with the current trend to perform multicentre studies. Moreover, manual labelling is time-consuming, so that most of the datasets may contain only a restricted number of annotated data. One way to handle heterogeneous data is to learn a unique classifier from the pooled

K.K. Bhatia and H. Lombaert (Eds.): MLMMI 2015, LNCS 9487, pp. 74–82, 2015.
DOI: 10.1007/978-3-319-27929-9_8

databases [1]. This, however, is likely to violate the standard supervised learning assumption that the training and test data come from the same statistical distribution and may lead to poor classification performance. One promising alternative is to investigate how transfer learning can adapt to this problem, specifically regarding two scenarios encountered in the clinical practice. The first situation happens when one wants to classify new target data originating from a different imaging protocol than that used to generate the original CAD training database. In that case, this original training database, referred to as the source domain, contains a large number of annotated samples, while the new dataset, referred to as the target domain, is composed of limited annotated samples. This situation directly fits the methodological framework of transductive transfer learning. The second case assumes that each database (from the source or target domain) contains a sufficient number of annotated data to build a separate classifier, but one wants to efficiently learn across these two domains. The question is how to boost the diagnostic performance that would be achieved by any of the CAD scheme trained on each domain separately, or on the pooled source and target databases. In this paper, we address the second case study by proposing two novel contributions in the field of subspace alignment methods that are part of domain adaptation, and comparing their performance with that of Adaptive SVM [2]. The medical context is that of prostate cancer screening based on multiparametric magnetic resonance imaging (mp-MRI). We recently achieved promising results with a CAD system that generates probability maps of malignancy based on the combination of a series of statistical, structural and functional features extracted from three MR sequences (T2, ADC and DCE) and an SVM classifier [3,4]. This system was trained over a clinical database consisting of 35 patient mpMRI exams acquired on a 1.5 T MR scanner. This scanner was recently replaced by a 3T machine, so that the patients database now aggregates 22 exams produced with a similar imaging and annotation protocol but on a 3T scanner.

Transfer learning has received a recent but increasing interest in the medical imaging community, for automated segmentation of MR neuroimaging [5], electron microscopy data [6], or classification of multimodality neuroimaging data [7,8]. As far as we know, these methods have not been applied yet to the challenging question of adapting labeled source and target domains to boost CAD diagnostic performance. This paper is organized as follows: Sect. 2 gives some background knowledge on transfer learning as well as recent developments in the field of subspace alignment methods. The two novel contributions are described in Sect. 3. Section 4 presents the experiments that were carried out on the two MRI databases to evaluate the achievable diagnostic performance based on these methods compared with that achieved without adaptation or with the Adaptive SVM. The results and discussion are given in the last section.

2 Background

Domain adaptation is a specific part of transfer learning in which we have two domains with an underlying distribution mismatch but lying in the same feature

space and corresponding to the same learning task. In the context of domain adaptation, there are two main settings, the unsupervised domain adaptation in which we only have access to big amount of labeled data from the source domain (the different distribution) but no labeled examples from the target domain. The other setting, is the semi supervised domain adaptation where there exists few labeled samples from the target domain in addition to the labeled samples of the source domain.There are different approaches to tackle this semi supervised adaptation problem, the simplest approach is the feature augmentation where each feature in the original space is mapped to an augmented space by duplicating the feature vector [9]. Another approach is the manifold learning that aims at finding a lower dimensional latent space and then aligns the two embedding subspaces [10]. A third approach represents the SVM based adaptation methods by either learning a target classifier that is close to the previously learned source classifier [2] or by training the SVM on all the samples from the two domains while giving less weights to those from the source domain [11]. In this work, we introduce a new simple approach that is based on finding a mapping of the source distribution to the target distribution in a semi supervised manner. We also suggest an extension to state of the art method on unsupervised domain adaptation called landmarks based subspace alignment [12] by making use of the available labeled examples from the target through the selection procedure of the landmarks. Furthermore, we compare these two methods with the Adaptive SVM [2] that has been used recently in the context of medical images [5] and shows a good performance. In the following, we explain each of the three approaches and finalize by comparing their performance on MRI prostate cancer dataset.

3 Methods

In the following methods, we deal with the case of domain adaptation in which we have a source dataset drawn from a distribution D_S and target dataset drawn from a different distribution D_T. The aim is to find a classifier with a low classification error on the target domain.

3.1 Adaptive SVM

The main idea is to learn independently an SVM classifier on the source distribution D_S and another SVM classifier on the available labeled samples from the target distribution D_T meanwhile constraining the new classifier to be as close as possible to the source classifier and in the same time minimizing the classification error on the target samples as proposed by Yang et al. [2]. The method adapts the classifier $f_S(x)$ learnt on D_S by adding a delta function $\triangle f(x)$:

$$f(x) = f_S(x) + \triangle f(x) = f_S(x) + w'x \tag{1}$$

In order to learn w, the method minimizes the following objective function:

$$\begin{aligned} min_w \frac{1}{2} \parallel w \parallel^2 + C \sum_{i=1}^{N} \xi_i \\ s.t.\ y_i f_S(x_i) + y_i w' x \geq 1 - \xi_i \\ \xi_i \geq 0 \\ \forall (x_i, y_i) \in D_T \end{aligned} \tag{2}$$

Where N is the number of the available target samples. Here the slack variables ξ_i have the same role as in regular SVM that is to minimize the classification error on the target samples. Distinctively, w penalizes the deviation of the new classifier $f(x)$ from the classifier $f_S(x)$ learnt on the source domain.

3.2 Proposed Approach: Source Scaling

Here, we aim at discovering a different simple approach in which we exploit possible transformation that puts the source samples as close as possible to the available target samples. This step will give the later learned classifier the ability to gain possible knowledge from the source dataset combined with the available target set. Our hypothesis is that the target samples are stemmed from a feature distribution that has been non uniformly scaled as compared to the source distribution. This might happen when source and target data originate from different scanners but almost similar imaging protocols. So, we want to find a mapping function that moves the source samples closer to the target samples with respect to their underlying labels. Suppose that we have a source dataset $S = \{x^p_{S_1}, .., x^p_{S_{Kp}}\} \cup \{x^n_{S_1}, .., x^n_{S_{Kn}}\}$ composed of positive and negative examples drawn from a marginal distribution D_S over X. We also have access to samples from the target dataset $T = \{x^p_{T_1}, .., x^p_{T_{Lp}}\} \cup \{x^n_{T_1}, .., x^n_{T_{Ln}}\}$ composed of positive and negative samples drawn from the marginal distribution D_T over X. We want to find the vectors $\alpha = \{\alpha_1, .., \alpha_F\}$ and $\beta = \{\beta_1, .., \beta_F\}$ (F is number of features) that minimize the following equations:

$$min_\alpha \parallel \frac{1}{K_p} \sum_{i=1}^{K_p} \alpha \circ x^p_{S_i} - \frac{1}{L_p} \sum_{j=1}^{L_p} x^p_{T_j} \parallel_2 \tag{3}$$

$$min_\beta \parallel \frac{1}{K_n} \sum_{i=1}^{K_n} \beta \circ x^n_{S_i} - \frac{1}{L_n} \sum_{j=1}^{L_n} x^n_{T_j} \parallel_2 \tag{4}$$

Where $\circ$ is the Hadamard product (element wise product), K_p and K_n are the number of positive and negative samples in the source dataset. Similarly, L_p and L_n are the number of positive and negative available samples in the target dataset. These equations have a closed form solution which is a scaling factor that matches the mean of the source distribution and target distribution regarding

each feature and class label. After obtaining the scaling vectors α and β, we map the source distributions to a new adapted one in order to minimize the distributions divergence.

$$\begin{aligned} Source\,Adapted\,Dataset = \; & A \circ X_S^p \cup B \circ X_S^n \\ Where\, A = \; & l^T \alpha \\ B = \; & l^T \beta \\ l = [1, .., 1]\, of\, length\, F & \end{aligned} \tag{5}$$

$X_S^p \in \mathbb{R}^{K_p \times F}$ and $X_S^n \in \mathbb{R}^{K_n \times F}$ denote the source positive samples and source negative samples respectively. The features are normalized after while to have a zero mean and standard deviation equals to one. Then, we can pool all the samples from the target and the adapted source together and learn any type of classifier. This scaling can be easily extended to any number of classes. Additionally, we can run a feature selection step [13] in order to eliminate some features that may exhibit a large difference and thus not suitable for adaptation.

3.3 Landmarks Based Approach

A third approach will be to visit the source and target samples looking for possible key points (landmarks) that will maximize the overlapping between the source and target distributions. This idea has been recently presented by Aljundi et al. [12] but in unsupervised manner. Here we show how we can use the same idea in semi supervised manner taking advantage of available labeled target samples. We first present the unsupervised selection of the landmarks. Having source samples X_S drawn for the source distribution D_S and target samples X_T drawn from the target distribution D_T, the method looks for possible candidates in both datasets. For each candidate point c, a mapping of the source data to a nonlinear feature Φ is performed using a Gaussian kernel:

$$\Phi_S(x_{S_i}, c) = exp(- \parallel x_{S_i} - c \parallel^2 /2s^2) \tag{6}$$

and similarly for the target points:

$$\Phi_T(x_{T_j}, c) = exp(- \parallel x_{T_j} - c \parallel^2 /2s^2) \tag{7}$$

Where s is the width of the kernel.Then, to compute the degree of overlapping between the two sets Φ_S and Φ_T, the two distributions are approximated as normal distributions and summarized by their means and standard deviations μ_S, σ_S, μ_T, σ_T. Here the overlapping is estimated by the integral of the dots products between the two distributions:

$$\int N(x \mid \mu_S, \sigma_S^2) N(x \mid \mu_T, \sigma_T^2) dx = N(\mu_S - \mu_T \mid 0, \sigma_{sum}^2) \tag{8}$$

$$overlap(\mu_S, \sigma_S; \mu_T, \sigma_T) = \frac{N(\mu_S - \mu_T \mid 0, \sigma_{sum}^2)}{N(0 \mid 0, \sigma_{sum}^2)} \tag{9}$$

Where σ^2_{sum} is the sum of σ^2_T and σ^2_S. The candidates that give an overlapping value above a threshold are selected. As the width of the kernel used for choosing the landmarks is highly related to the candidate itself, the method follows a multi scale fashion by looping among different values for the kernel width stemmed from the percentiles of the euclidian distances between the two datasets. The selected landmarks are then used to project the source and target distributions in a shared space using a Gaussian kernel. In the original method, there is an additional subspace alignment step that is performed after while. Here, we focus on the landmarks selection part that can be extended to the supervised manner.

Supervised Landmarks Selection: Since, we have access to some labeled target samples, we now show how we can plug this information in the landmarks selection procedure:
In a similar manner, we loop over the source and target points looking for possible candidates and map the source and target data to nonlinear feature $\Phi_S(x_{S_i}, c)$ and $\Phi_T(x_{T_j}, c)$ using a Gaussian kernel. Instead of using one normal distribution to model the source data and target data, we model the positive source data and negative source data independently $N(x \mid \mu^p_S, \sigma^p_S)$, $N(x \mid \mu^n_S, \sigma^n_S)$ and similarly for the positive and negative target samples $N(x \mid \mu^p_T, \sigma^p_T))$, $N(x \mid \mu^n_T, \sigma^n_T))$. We estimate the overlapping values between the following pairs: {(target positive and source positive), (target negative and source negative), (target positive and target negative), (target positive and source negative) and (target negative and source positive)}. Then, we use a weighted sum over the overlapping between the pairs of the same class labels minus the overlapping between pairs of different labels. The final overlapping value for a given candidate is:

$$\begin{aligned} TotalOverlap = \; & 1/2(overlap(\mu^p_T, \sigma^p_T; \mu^p_S, \sigma^p_S) + overlap(\mu^n_T, \sigma^n_T; \mu^n_S, \sigma^n_S)) \\ & -1/3(overlap(\mu^p_T, \sigma^p_T; \mu^n_T, \sigma^n_T) + overlap(\mu^p_T, \sigma^p_T; \mu^n_S, \sigma^n_S) \\ & +overlap(\mu^n_T, \sigma^n_T; \mu^p_S, \sigma^p_S)) \end{aligned} \quad (10)$$

Then, as in the original method we select the candidates that give an overlapping level above a threshold fixed experimentally. Notice that Eq. (10) can be easily adapted to case of multiple classes.

4 Experiments

A series of experiments was carried out in order to compare the different discussed approaches in the context of semi supervised domain adaptation. The source database consisted of 35 patients who underwent mp-MR imaging on a 1.5T clinical MR scanner (Symphony, Siemens Medical Systems, Germany) following the protocol described in [4]. Each tumor or suspicious tissue was outlined by an expert radiologist over the three MR sequences. The nature of the tumor as well as its Gleason score characterizing cancer aggressiveness were confirmed by an anatomopathologist. The target database consisted of 22 patients who underwent mp-MR imaging on a 3T clinical MR scanner (Discovery MR750

Table 1. Comparison of performance between training on target and training on both source and target regarding the three target sets

Method	2000 D_T Samples	5000 D_T Samples	10000 D_T Samples
Only Target	0.722	0.734	0.737
Target and Source (NA)	0.727	0.735	0.741

Table 2. Performance of the compared methods on different adaptation settings

Method	D_S+2000 D_T Samples	D_S +5000 D_T Samples	D_S +10000 D_T Samples
NA	0.72	0.73	0.74
ASVM	0.71	0.73	0.72
SLA	0.76	0.75	0.75
SC	0.73	0.75	0.76

General Electric Medical Systems, USA) following similar acquisition protocol described in [14]. We want to show the adaptation performance and the possible boosting in performance on 3T dataset by using the 1.5T dataset. We first randomly drew 10000 voxels out of the 1.5T data, balancing the number of samples among the different classes of tissues (normal and cancer lesions of different GL score). A similar random sampling was performed on the 3T database to produce 3 different subsets, of 2000, 5000 and 10000 samples respectively. Three different training databases were constructed by pooling the sampled source dataset with each of the subsampled target sets. The test dataset was the biggest target set which is composed of 10000 samples. These training and test databases served to compare the three previously explained approaches which are Adaptive SVM (ASVM), Supervised Landmarks approach (SLA) and Source Scaling approach (SC) in addition to the baseline method consisting in combining the samples from the two different distributions without any adaptation (NA). Cross validation was used to perform both the adaptation step and the classification with a linear SVM within the same loop of the leave-one-patient out (LOPO) strategy. The area under the roc curve, AUC, was used as the performance metric, which is most suitable in our case due to the unbalancing in the data (The percent of cancer voxels to the normal tissues). Figure 1 shows, for the different methods, an example probability map of cancer overlaid a 3T MR T2 transverse of a patient with two aggressive cancer lesions (GL = 8) in the peripheral zone. The voxels detected by the methods as normal tissues are shown as transparent. Tables 1 and 2 show the detection performance achieved with the different methods and considering the different training datasets.

4.1 Results Analysis

Table 1 shows that pooling the source dataset without adaptation doesnt improve the performance over the use of only the target set. On the other hand, as shown

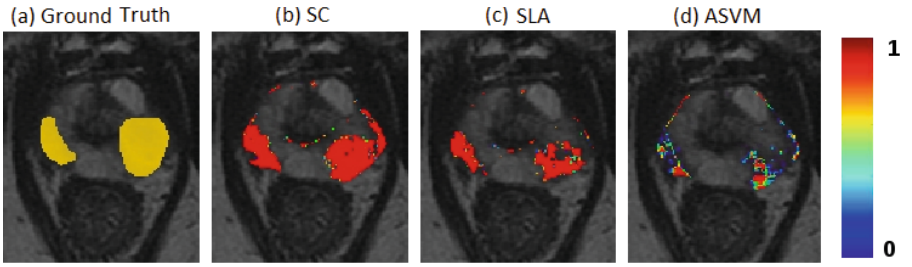


Fig. 1. Example transverse slice of a patient with two aggressive cancer lesions. (a) ground truth pixel labelling of the two lesions, (b) probability map produced by SC method, (c) probability map produced by SLA method, (d) probability map produced by ASVM.

in Table 2, the use of domain adaptation methods SLA and SC improved the raw performance over the different sets of samples. However, the use of Adaptive SVM did not show any boosting in performance which is explained by the fact that such a method better handles the case of having access to few annotated target samples which is not the aim of our study. This is mainly because the Adaptive SVM constrains the target learned classifier to be close to a classifier learned on the source set, which might be appropriate in the case of having an insufficient number of target samples to learn even a moderate classifier. It should be noticed also that the Source Scaling (SC) method gives good results when having access to reasonable amount of target samples and is able to transfer the knowledge from the source samples to further improve the target classifier performance. In that case,despite its simplicity and its ability of keeping the raw features which is important in some medical situations, the SC method gives similar results to the state of the art method that we extend to the semi supervised setting (SLA).It is worth noting that this performance is able to be increased up to the AUC = 0.78 by adding a feature selection step [13], which is competitive to the state of the art CAD system [3]. When considering small target samples ($D_T = 2000$) the SLA method is shown to overcome the SC performance.

5 Conclusion

We present novel contributions in the field of domain adaptation for medical imaging. Our aim is to boost the performance of computer aided diagnostic system by efficiently combining heterogeneous prostate MR imaging databases. We suggest a simple approach based on scaling the features of the source set to the target set. This transformation is likely to be suitable for medical imaging cases where the shift is stemmed from using different types of scanner that imposes different resolution or different contrast. The performance of this approach based on matching the mean of the source and target distribution increases by gaining access to more target samples. We also suggest an extension to a state of the

art method on unsupervised domain adaptation by using available target labels which seems to give stable performance even when using the smallest target set. As a future work, we will test the performance of the supervised landmark based method on the target datasets of lower size. We also plan to run a large series of experiments and examine the adaptation performance with respect to the both available datasets and explore a possible solution to work on patients from different scanners.

References

1. Hong, S.J.J., Kim, H., Schrader, D., Bernasconi, N., Bernhardt, B.C., Bernasconi, A.: Automated detection of cortical dysplasia type II in MRI-negative epilepsy. Neurology **83**(1), 48–55 (2014)
2. Yang, J., Yan, R., Hauptmann, A.G.: Cross-domain video concept detection using adaptive svms. In: Proceedings of the 15th International Conference on Multimedia, Multimedia 2007, pp. 188–197. ACM, New York (2007)
3. Lehaire, J., Flamary, R., Rouviere, O., Lartizien, C.: Computer-aided diagnostic system for prostate cancer detection and characterization combining learned dictionaries and supervised classification. In: 2014 IEEE International Conference on Image Processing (ICIP), pp. 2251–2255. IEEE, October 2014
4. Niaf, E., Rouvière, O., Mège-Lechevallier, F., Bratan, F., Lartizien, C.: Computer-aided diagnosis of prostate cancer in the peripheral zone using multiparametric MRI. Phys. Med. Biol. **57**(12), 3833–3851 (2012)
5. van Opbroek, A., Ikram, M.A., Vernooij, M.W., de Bruijne, M.: Transfer learning improves supervised image segmentation across imaging protocols. IEEE Trans. Med. Imaging **34**(5), 1018–1030 (2015)
6. Becker, C., Christoudias, C.M., Fua, P.: Domain adaptation for microscopy imaging. IEEE Trans. Med. Imaging **34**(5), 1125–1139 (2015)
7. Guerrero, R., Ledig, C., Rueckert, D.: Manifold alignment and transfer learning for classification of Alzheimer's disease. In: Wu, G., Zhang, D., Zhou, L. (eds.) MLMI 2014. LNCS, vol. 8679, pp. 77–84. Springer, Heidelberg (2014)
8. Jie, B., Zhang, D., Cheng, B., Shen, D.: Manifold regularized multi-task feature selection for multi-modality classification in Alzheimer's disease. In: Mori, K., Sakuma, I., Sato, Y., Barillot, C., Navab, N. (eds.) MICCAI 2013, Part I. LNCS, vol. 8149, pp. 275–283. Springer, Heidelberg (2013)
9. Daumé, H.: Frustratingly easy domain adaptation. CoRR arXiv preprint arXiv:0907.1815 (2009)
10. Wang, C.: A geometric framework for transfer learning using manifold alignment. Ph.D. thesis (2010)
11. Wu, P., Dietterich, T.G.: Improving svm accuracy by training on auxiliary data sources. In: Proceedings of the Twenty-first International Conference on Machine Learning, ICML 2004, p. 110. ACM, New York (2004)
12. Aljundi, R., Emonet, R., Muselet, D., Sebban, M.: Landmarks-based kernelized subspace alignment for unsupervised domain adaptation, June 2015
13. Rakotomamonjy, A.: Variable selection using svm based criteria. J. Mach. Learn. Res. **3**, 1357–1370 (2003)
14. Bratan, F., Niaf, E., Melodelima, C., Chesnais, A.L.L., Souchon, R., Mège-Lechevallier, F., Colombel, M., Rouvière, O.: Influence of imaging and histological factors on prostate cancer detection and localisation on multiparametric MRI: a prospective study. Eur. Radiol. **23**(7), 2019–2029 (2013)

Segmentation

Feature-Space Transformation Improves Supervised Segmentation Across Scanners

Annegreet van Opbroek[1(✉)], Hakim C. Achterberg[1], and Marleen de Bruijne[1,2]

[1] Biomedical Imaging Group Rotterdam, Departments of Medical Informatics and Radiology, Erasmus MC - University Medical Center Rotterdam, Rotterdam, The Netherlands
a.vanopbroek@erasmusmc.nl
[2] Department of Computer Science, University of Copenhagen, Copenhagen, Denmark

Abstract. Image-segmentation techniques based on supervised classification generally perform well on the condition that training and test samples have the same feature distribution. However, if training and test images are acquired with different scanners or scanning parameters, their feature distributions can be very different, which can hurt the performance of such techniques.

We propose a feature-space-transformation method to overcome these differences in feature distributions. Our method learns a mapping of the feature values of training voxels to values observed in images from the test scanner. This transformation is learned from unlabeled images of subjects scanned on both the training scanner and the test scanner.

We evaluated our method on hippocampus segmentation on 27 images of the Harmonized Hippocampal Protocol (HarP), a heterogeneous dataset consisting of 1.5T and 3T MR images. The results showed that our feature space transformation improved the Dice overlap of segmentations obtained with an SVM classifier from 0.36 to 0.85 when only 10 atlases were used and from 0.79 to 0.85 when around 100 atlases were used.

Keywords: Brain · Hippocampus · Machine learning · MRI · Transfer learning

1 Introduction

Supervised voxelwise classification, where manually labeled images are used to train a voxel classifier is a very popular approach for many medical-image segmentation tasks. For these methods to perform optimally they need to be trained on images that are representative of the test images. Such training images may however not always be available, since training and test images are only truly representative if they are acquired with the same scanner and scanning parameters and concern similar patient groups. Differences in scanners and patient

K.K. Bhatia and H. Lombaert (Eds.): MLMMI 2015, LNCS 9487, pp. 85–93, 2015.
DOI: 10.1007/978-3-319-27929-9_9

groups can result in differences in the voxels' feature distributions, which is likely to deteriorate a classifier's performance.

A way to tackle this problem is by applying transfer learning [1]. So far, most transfer-learning methods used in medical image segmentation concern transfer classifiers, which are usually based on some form of sample weighting [2]. In this paper we propose a different, complementary, transfer-learning approach. Our method aims at finding a feature-space transformation (FST) between training images acquired with one scanner and a test image from a different scanner, based on unlabeled images from one or multiple subjects scanned on both scanners. This FST is used to map the feature values of training voxels to values observed in voxels from the test scanner.

Our method is somewhat similar to image synthesis methods, which transform image intensities of images acquired with a certain scanner and contrast to values observed in images that were scanned with a different scanner or contrast. Roy et al. [3] for example, split up a source image into patches, which are all matched to patches of an atlas that is scanned with both scanners/contrasts. The corresponding patches from the other scanner/contrast are then combined into a final image. Such a method could be used to first transform intensities of training images, after which features can be extracted and a classifier can be trained. Image synthesis methods are however likely to produce images that are either slightly noisy of slight smoothed compared to the representation of test images. Such differences can lead to very big differences in feature representations between training and test data. In this paper we therefore investigate whether it is beneficial to directly transform all features.

We evaluated whether the proposed FST can improve performance of a voxel classifier for hippocampus segmentation in MR images. The most common approach to hippocampus segmentation is by multi-atlas registration [4]. One of many label-fusion techniques can then be used to combine the atlases into a final segmentation. [5] provides an overview and comparison of various approaches. Machine learning for hippocampus segmentation has gained attention in the previous years. Coupé et al. [6] performed voxelwise classification based on a patch surrounding the voxel. Classification is performed by a similarity-weighted voting of nearest-neighbor training patches in feature space. Powell et al. [7] and Van der Lijn et al. [8] showed that combining multi-atlas registration and voxelwise classification can improve performance compared to multi-atlas registration alone.

In this paper we extend the work of [7,8] on combining multi-atlas registration with voxel classifiers. We investigated whether the presented FST allows for such classifiers to be trained on images from different scanners than the test scanner.

2 Methods

2.1 Notation

Let $\boldsymbol{x}_i^s \in \mathbb{R}^d$ denote sample (voxel) i in a training image from scanner s and $y_i^s \in \{0, 1\}$ its label. $\boldsymbol{x}_i^s$ contains a value for each of the d measured features.

All training images from s together provide a set of N_s training samples and corresponding labels $\{\boldsymbol{x}_i^s, y_i^s\}_{i=1}^{N_s}$, which have a d-dimensional feature distribution F_s. A test image from scanner t gives a set of N_t test samples with unknown labels, $\{\boldsymbol{x}_i^t\}_{i=1}^{N_t}$ from feature distribution F_t. If samples originate from different scanners we assume their feature distributions to be different ($F_t \neq F_s$). The goal of our method is to learn a feature-space transformation (FST) from the training scanner s to the test scanner t: $f_{s\to t} : F_s \to F_t$. Once we know $f_{s\to t}$ we can use it to transform the training samples $\boldsymbol{x}_i^s$ from F_s to F_t by transforming it to $\boldsymbol{x}_i^s + f_{s\to t}(\boldsymbol{x}_i^s)$.

2.2 Determining Voxel-To-Voxel Correspondence

The FST is learned from one or multiple subjects that are scanned with both s and t. We will call these two images of the same subject the source image and the target image, which together form a source-target pair. These pairs should be acquired within a short time interval from each other and no segmentation labels are required for these images. The source images provide $\tilde{N}_s$ source voxels $\tilde{\boldsymbol{x}}_i^s$ and the target images provide $\tilde{N}_t$ target voxels $\tilde{\boldsymbol{x}}_i^t$. We perform an affine registration with a nearest-neighbor interpolation of the target images to their corresponding source images. This provides us with a voxelwise correspondence for every sample $\tilde{\boldsymbol{x}}_i^s$ to a sample $\tilde{\boldsymbol{x}}_\ell^t$:

$$\forall i : \ \exists\, \ell : \tilde{\boldsymbol{x}}_i^s \to \tilde{\boldsymbol{x}}_\ell^t \,. \tag{1}$$

2.3 Transforming Training Samples

Next, the training samples $\boldsymbol{x}_i^s$ are mapped from F_s to F_t based on the voxel pairs in Eq. 1. For each training sample $\boldsymbol{x}_i^s$ we determine the closest k source samples $\{\tilde{\boldsymbol{x}}_{m_1}^s, \tilde{\boldsymbol{x}}_{m_2}^s, \ldots, \tilde{\boldsymbol{x}}_{m_k}^s\}$ in feature space. The FST of $\boldsymbol{x}_i^s$ equals the median[1] difference of these k source samples to their corresponding target samples:

$$f_{s_j\to t}(\boldsymbol{x}_i^s) = \text{median}(\tilde{\boldsymbol{x}}_{\ell_1}^t - \tilde{\boldsymbol{x}}_{m_1}^s, \tilde{\boldsymbol{x}}_{\ell_2}^t - \tilde{\boldsymbol{x}}_{m_2}^s, \ldots, \tilde{\boldsymbol{x}}_{\ell_k}^t - \tilde{\boldsymbol{x}}_{m_k}^s), \tag{2}$$

where $\tilde{\boldsymbol{x}}_{\ell_n}^t$ $(n = 1, 2, \ldots, k)$ is the voxel pair of $\tilde{\boldsymbol{x}}_{m_n}^s$ defined in Eq. 1. We used the median to make sure that the chosen transformation is one that is observed in the pairs, unlike the mean, which could result in implausible transformations.

Increasing k increases the regularization, which results in a smoother transformation. By performing this regularization in feature space image details such as tissue edges are left in tact.

Note that the formulation is easily expendable to multiple source scanners by individually transforming each of them to the target distribution.

[1] We used the robust median, which gives the point that has minimal total distance to all other points.

3 Experiments

Data Description. We applied our method to hippocampus segmentation on 27 images of the HarP dataset [9]. We used the preliminary release, which consist of 100 images with manual hippocampus segmentations. All images are T1-weighted MR images, acquired with either a 1.5T or a 3T scanner. The images were scanned at a total of 31 sites with different scanners (Philips, GE, and Siemens scanners, various scanner types). See Fig. 2 for an example of differences between images from a 1.5T and a 3T scanner.

We segmented a subset of 27 of these 100 images, which consisted of all images that were acquired at one of the six sites that acquired both 1.5T and 3T images for the HarP dataset. We trained on images acquired with the 1.5T scanner and tested on images acquired at the same site with the 3T scanner and vice versa. The source-target pairs consisted of (unlabeled) ADNI [10] images (which are not in the HarP dataset). These images concern subjects scanned within 30 days with both the 1.5T and the 3T scanner of the site.

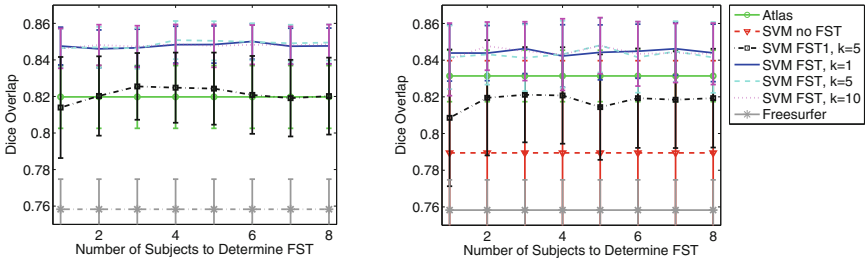
Multi-Atlas Hippocampus Probability. For each test sample (voxel) we obtained an atlas probability by non-rigid registration of images of the HarP dataset. We did experiments with (1) only 10 randomly chosen images as atlas and (2) all images from other scanners than the test image as atlas (this resulted in 94 to 99 atlases, depending on how many of the 100 images were scanned with the test scanner). Each of the atlas images was registered to the test image, after which the obtained transformation was applied to the binary manual segmentations. Averaging over all atlases gives an atlas probability, which was used as a feature in the voxel classifier. For the training set this feature was obtained by registration of the atlases to the training images. The classifier was trained and tested on a region of interest (ROI) consisting of all voxels with a non-zero atlas probability.

Registrations. All registrations were performed with Elastix [11] based on maximizing mutual information within a brain mask. These masks were generated with the brain-extraction tool (BET) [12]. We used the registration settings of [13], which were visually optimized for ADNI data. The source-target pairs were first rigidly and then affinely registered to each other, the HarP images were subsequently registered rigidly, affinely, and non-rigidly.

Features. We used a total of 11 features: 10 image-intensity features and the atlas-probability feature. The intensity features consisted of a subset of the features of [8]: (1) the original T1 intensity, (2–4) the intensity after a Gaussian smoothing with $\sigma = 1, 2, 4\,\mathrm{mm}^3$, (5–10) the Gradient Magnitude and Laplacian after convolution with a Gaussian kernel with $\sigma = 1, 2, 4\,\mathrm{mm}^3$. Before calculating the features we performed image normalization with a 4th-96th percentile range matching within the brain mask. Only the 10 image-intensity features were transformed with the FST. For each image all features were normalized to zero mean and unit variance within the whole image, except for the atlas feature, which was normalized within the ROI.

Classifier and Parameters. Voxelwise classification was performed with a support vector machine (SVM) with a Gaussian kernel. Separate classifiers were trained for the left and right hippocampus. For each site the SVM parameter C and the kernel parameter γ were set with a cross-validation experiment on the images of the five training sites. All SVM classifiers were trained on a total of 10 000 random training samples within the ROI from all training images. The FSTs were determined on all training voxels within the brain mask. We experimented with multiple values for the FST parameter k ($k = 1, 5, 10, 50, 100$) and with multiple numbers of source-target pairs (1 to 8).

Compared Methods. We compared the performance of: (1) *Atlas*: multi-atlas segmentation, which was obtained by thresholding the atlas probability map at 0.5, (2) *SVM no FST*: an SVM classifier without the FST, (3) *SVM FST1*: an SVM where the T1 intensity was transformed with the FST and the other intensity features were derived from the resulting image, (4) *SVM FST*: the proposed method: an SVM where all intensity features were transformed with the FST, and (5) Freesurfer[2] [14] version 5.1.0, a state-of-the-art brain-structure-segmentation tool. The performance of the methods was measured in terms of the Dice overlap between the manual and the automatic segmentation.



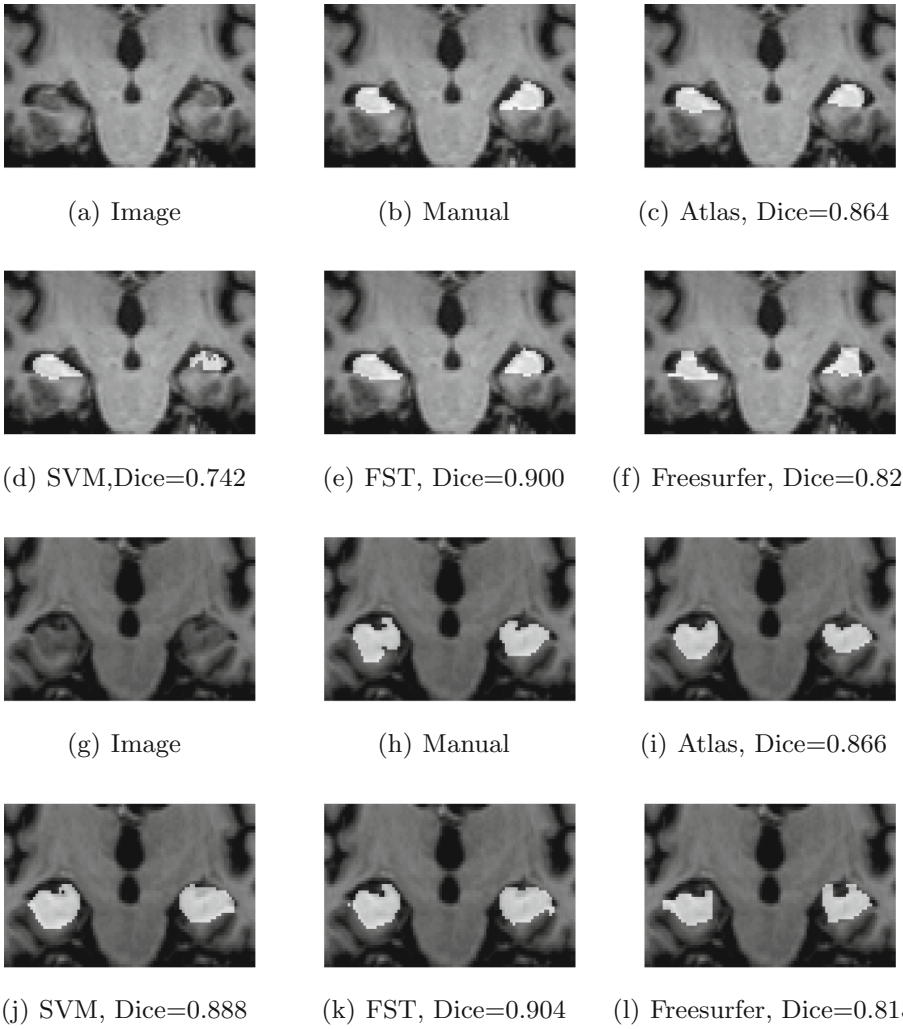
(a) 10 Atlases (b) All Atlases

Fig. 1. Mean dice overlap and 95 %-confidence interval of the mean for (1) multi-atlas segmentation, (2) SVM without FST, (3) SVM with the FST only on the intensity with $k = 5$, (4–6) SVM with the FST on all features for $k = 1, k = 5, k = 10$, and (7) Freesurfer. (a) gives the result for 10 atlases, (b) for all atlases, both as a function of the number of source-target pairs. For (a) *SVM no FST* gave a mean Dice of 0.36.

4 Results

Figure 1(a) shows the performance of the five methods when 10 atlases were used, and Fig. 1(b) when all atlases were used, both as a function of the number of source-target pairs used for the FST. In both experiments our SVM with FST performed best. For 10 atlases the SVM without FST performed poorly with

[2] Documented and freely available at http://surfer.nmr.mgh.harvard.edu/.



(a) Image (b) Manual (c) Atlas, Dice=0.864

(d) SVM,Dice=0.742 (e) FST, Dice=0.900 (f) Freesurfer, Dice=0.823

(g) Image (h) Manual (i) Atlas, Dice=0.866

(j) SVM, Dice=0.888 (k) FST, Dice=0.904 (l) Freesurfer, Dice=0.815

Fig. 2. Example of obtained segmentations for *Atlas*, *SVM no FST*, *SVM FST* ($k = 1$, 1 source-target pair), and Freesurfer. (a)-(f) show the results for a 3T image, (g)-(l) for a 1.5T image. The Dice scores for the slice are shown under the images.

a mean dice of and 0.36. Freesurfer also performed rather poorly with a mean dice of 0.76. Multi-atlas segmentation performed well, and applying the FST only to the intensity performed similarly; both obtained a mean dice of 0.82. With a mean dice of 0.85 our method significantly outperformed all four other methods[3]. When all atlases were used the multi-atlas segmentation increased

[3] $p = 10^{-6}$, $p = 10^{-9}$, $p = 10^{-5}$, and $p = 10^{-3} \sim 10^{-5}$ for *SVM no FST*, Freesurfer, *Atlas*, and *SVM FST1* respectively with a two-sided paired t-test.

performance to 0.83, the SVM without FST increased performance to 0.79, and the SVM FST1 performed similarly as with 10 atlases (Dice≈0.82). Our SVM with FST still performed best with a mean dice of 0.84 to 0.85 depending on the k and the number of source-target pairs and significantly better than all methods except from the multi-atlas method[4].

Contrary to the other two methods, SVM FST and SVM FST1 decreased in performance when all atlases were used compared to when 10 atlases were used. This is because the SVMs classified all voxels with an atlas probability above zero. When more atlases are used it becomes more likely that some poorly registered atlases are included and the number of test voxels increases, resulting in an increase in potential false positives. This problem could be solved by reducing the size of the ROI, e.g. by setting a higher threshold on the atlas probability.

The figures show a very small increase in the performance of the SVM FST when the number of source-target pairs increases, but there were no significant differences. There were no significant differences between the tried values for the regularization parameter k. We also experimented with $k = 50$ and $k = 100$ (not shown here), which slightly decreased the performance. SVM FST1 performed slightly better for $k = 5$ (shown in Fig. 1) than for $k = 1$ and $k = 10$.

Figure 2 shows the segmentations for two of the 27 subjects. Note that the Dice scores for the shown slice are higher that the scores in Fig. 1 because the shown slice has a relatively large fraction of hippocampus voxels. For the SVM without FST on 10 atlases the shown segmentation is much better than average.

5 Conclusion and Discussion

We presented a feature-space transformation (FST) method to cope with training images that are obtained with different scanners or scanning protocols than a test image. Our FST maps the training voxels to the feature distribution of the test voxels based on images of one or a few subjects that are scanned on both the training and the test scanner. After application of this FST any classifier could be used for classification.

Experiments on hippocampus segmentation in a multi-center dataset showed that our FST can greatly improve classification results. In experiments with a subset of 10 atlases it increased the mean Dice score of an SVM classifier from 0.36 to 0.85. In experiments with around 100 atlases our FST increased the Dice overlap from 0.79 to 0.85, and this result can be expected to improve further by decreasing the ROI. The results also showed a significant improvement of mapping all features that are used for the classification over mapping only the intensity, which obtained a mean Dice of only 0.82 for both settings. Note that due to the small number of labeled images per site in this study the SVM was often trained on only 1 or 2 images. If more training images would be available per site we can expect the accuracy of our FST SVM to improve and reach values similar to those reported by other hippocampus-segmentation methods [4].

[4] $p = 10^{-3}$, $p = 10^{-6}$, $p = 0.1$, $p = 10^{-2} \sim 10^{-3}$ for *SVM no FST*, Freesurfer, *Atlas*, and *SVM FST1* respectively with a two-sided paired t-test.

The experiments also showed that the incorporation of an SVM classifier with our FST can significantly improve performance over multi-atlas segmentation, especially when few atlases are available. This is in line with previous work on brain-structure segmentation [7,8], which showed for single-center data that atlas-based segmentation can be improved by incorporation of a voxelwise classifier. Note that while only a simple average atlas probability was used, more extensive label-fusion techniques could be easily incorporated in the SVM. We also showed that our method significantly outperformed the readily available segmentation tool Freesurfer.

We experimented with regularization by non-local median filtering to smooth out possible noise in the FST due to e.g. image noise and mis registrations. Increasing the amount of regularization did not significantly improve the results. We also experimented with determining the FST on multiple subjects scanned on training and test scanner. For the FST only on the intensity this significantly improved the performance, but for the FST on all features this difference was not significant. That more regularization and incorporation of more subjects did not improve performance might be because of the good voxelwise correspondence within subjects after registration as well as a low noise level in these images. Whether it would also be possible to obtain the FST from the training and test images might be an interesting topic for further research. In this case the assumption of a voxelwise correspondence after registration might not hold, which means that regularization becomes much more important.

The presented method can be applied not only to hippocampus segmentation, but to classification in a wide range of applications where registration of two images from the same subject can give an FST. The presented method can be of particular interest for longitudinal studies, which often scan a few subjects with multiple scanners to check for reproducibility and scanning problems. We believe that since our method can construction an FST from unlabeled images of only a single subject our method can be very valuable in practice.

Acknowledgments. This research is financed by The Netherlands Organization for Scientific Research (NWO).

References

1. Pan, S., Yang, Q.: A survey on transfer learning. IEEE Trans. Knowl. Data Eng. **22**(10), 1345–1359 (2010)
2. Van Opbroek, A., Ikram, M., Vernooij, M., De Bruijne, M.: Transfer learning improves supervised image segmentation across imaging protocols. IEEE Trans. Med. Imaging **34**(5), 1018–1030 (2014)
3. Roy, S., Carass, A., Prince, J.: A compressed sensing approach for MR tissue contrast synthesis. In: Székely, G., Hahn, H.K. (eds.) IPMI 2011. LNCS, vol. 6801, pp. 371–383. Springer, Heidelberg (2011)
4. Dill, V., Franco, A., Pinho, M.: Automated methods for hippocampus segmentation: the evolution and a review of the state of the art. Neuroinformatics **1**, 1–18 (2014)

5. Artaechevarria, X., Munoz-Barrutia, A., Ortiz-de-Solórzano, C.: Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans. Med. Imaging **28**(8), 1266–1277 (2009)
6. Coupé, P., Manjón, J., Fonov, V., Pruessner, J., Robles, M., Collins, D.: Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. Neuroimage **54**(2), 940–954 (2011)
7. Powell, S., Magnotta, V., Johnson, H., Jammalamadaka, V., Pierson, R., Andreasen, N.: Registration and machine learning-based automated segmentation of subcortical and cerebellar brain structures. Neuroimage **39**(1), 238–247 (2008)
8. Van der Lijn, F., De Bruijne, M., Klein, S., Den Heijer, T., Hoogendam, Y., Van der Lugt, A., Breteler, M., Niessen, W.: Automated brain structure segmentation based on atlas registration and appearance models. IEEE Trans. Med. Imaging **31**(2), 276–286 (2012)
9. Boccardi, M., Bocchetta, M., Morency, F., Collins, D., Nishikawa, M., Ganzola, R., Grothe, M., Wolf, D., Redolfi, A., Pievani, M., et al.: Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. Alzheimer's Dement. **11**(2), 175–183 (2015)
10. Mueller, S., Weiner, M., Thal, L., Petersen, R., Jack, C., Jagust, W., Trojanowski, J., Toga, A., Beckett, L.: Ways toward an early diagnosis in Alzheimer's disease: the Alzheimer's disease neuroimaging initiative (ADNI). Alzheimer's Dement. **1**(1), 55–66 (2005)
11. Klein, S., Staring, M., Murphy, K., Viergever, M., Pluim, J.: Elastix: a toolbox for intensity-based medical image registration. IEEE Trans. Med. Imaging **29**(1), 196–205 (2010)
12. Smith, S.: Fast robust automated brain extraction. Hum. Brain Mapp. **17**(3), 143–155 (2002)
13. Bron, E., Steketee, R., Houston, G., Oliver, R., Achterberg, H., Loog, M., Swieten, J., Hammers, A., Niessen, W., Smits, M., et al.: Diagnostic classification of arterial spin labeling and structural MRI in presenile early stage dementia. Hum. Brain Mapp. **35**(9), 4916–4931 (2014)
14. Fischl, B., Salat, D., Busa, E., Albert, M., Dieterich, M., Haselgrove, C., Van der Kouwe, A., Killiany, R., Kennedy, D., Klaveness, S., et al.: Whole brain segmentation: automated labeling of neuroanatomical structures in the human brain. Neuron **33**(3), 341–355 (2002)

Discriminative Dimensionality Reduction for Patch-Based Label Fusion

Gerard Sanroma[1(✉)], Oualid M. Benkarim[1], Gemma Piella[1], Guorong Wu[2], Xiaofeng Zhu[2], Dinggang Shen[2], and Miguel Ángel González Ballester[1,3]

[1] Department of Information Technologies and Communication, UPF, Barcelona, Spain
gerard.sanroma@upf.edu
[2] Department of Radiology and BRIC, UNC-Chapel Hill, Chapel Hill, USA
[3] ICREA, Barcelona, Spain

Abstract. In this last decade, multiple-atlas segmentation (MAS) has emerged as a promising technique for medical image segmentation. In MAS, a novel target image is segmented by fusing the label maps of a set of annotated images (or atlases), after spatial normalization. Weighted voting is a well-known label fusion strategy consisting of computing each target label as a weighted average of the atlas labels in a local neighborhood. The weights, denoting the local anatomical similarity of the candidate atlases, are often approximated using image-patch similarity measurements. Such an approach, known as patch-based label fusion (PBLF), may fail to discriminate the anatomically relevant patches in challenging regions with high label variability. In order to overcome this limitation we propose a supervised method that embeds the original image patches onto a space that emphasizes the appearance characteristics that are critical for a correct labeling, while supressing the irrelevant ones. We show that PBLF using the embedded patches compares favourably with state-of-the-art methods in brain MR image segmentation experiments.

1 Introduction

Medical image segmentation aims at estimating a dense label map of the anatomical structures in medical images, such as magnetic resonance images (MRI) of the human brain. Quantitative analysis of segmentation data is useful in many fields such as the neurosciences, where the morphometric analysis of brain structures helps characterizing the progression of diseases such as Alzheimer and Schizophrenia [2]. Manual annotation is a tedious time-consuming process which has to be done by trained experts and thus, automatic methods are highly valuable.

Partly enhanced by the success of image registration, multiple-atlas segmentation (MAS) has recently gained attention for segmenting medical images. Three main steps are involved in MAS: (i) the *image registration* step registers each individual atlas onto the target image [7], (ii) the *atlas selection* step selects the best atlases for segmenting a particular target image [1,13,14], and

K.K. Bhatia and H. Lombaert (Eds.): MLMMI 2015, LNCS 9487, pp. 94–103, 2015.
DOI: 10.1007/978-3-319-27929-9_10

(ii) the *label fusion* step fuses the registered label maps onto a consensus segmentation [3,4,10–12,15–17,21,22].

By combining the labels from multiple atlases, the label fusion step can compensate for the registration errors by the individual atlases. Even a simple label fusion strategy such as majority voting [11] (which assigns each target voxel to the label appearing most frequently among the corresponding atlas labels) yields better segmentation performance than any of the single atlases used individually [10].

Another commonly used label fusion strategy is weighted voting, in which the label on each target point is computed as a weighted average of the atlas labels, where the weights reflect the estimated anatomical similarity between the target and each atlas. A critical issue here, is how to set the weights that accurately reflect the anatomical correspondence. One common approach, adopted in patch-based label fusion (PBLF), consists in estimating the weights based on the local similarity between the atlas and the target image patches [3,4,12,17,22].

However, there are uncertain regions, such as the interfaces between two anatomical structures, where very similar atlas patches may bear different labels. In such cases, the appearance cues responsible for a correct discrimination may be too weak to be correctly captured by simple image similarity measurements. In order to overcome this limitation we propose a method that learns an embedding of the image patches so that the relevant variations for a correct discrimination are emphasized while the misleading ones are supressed. We pose this problem in a supervised learning setting, where we seek the linear mapping of the image patches that simultaneously (i) maximizes the similarity of the target patch with its true neighbors (i.e., similar atlas patches with the same label) and (ii) minimizes the similarity with its false neighbors (i.e., similar atlas patches with different label).

The proposed method bears some similarity with manifold learning methods such as neighborhood preserving embeddings (NPE) [9] and locality preserving projections (LPP) [8] in the sense that it aims at preserving the true neighborhood but it simultaneously enforces an additional discriminative component aimed at simultaneously maximizing the separation between false neighbors.

The weights obtained by the proposed method using an example target patch are illustrated in Fig. 1. As we can see in the top-left plot, using the similarity in the original image space, a fair amount of atlas patches with the wrong label (denoted in red) accumulate a considerable amount of weight. On the other hand, using the similarity of the projected patches with the proposed method, the weights of the wrong atlas labels are considerably reduced, while still maintaining a significant amount of weight for the atlas patches with the correct labels (denoted in blue), as shown in the top-right plot.

The contributions of the proposed method are three-fold:

- The embeddings are learned offline in a common space so that they can be readily applied to any new target image.
- The learned embeddings can be plugged into any existing PBLF method to enhance its performance.

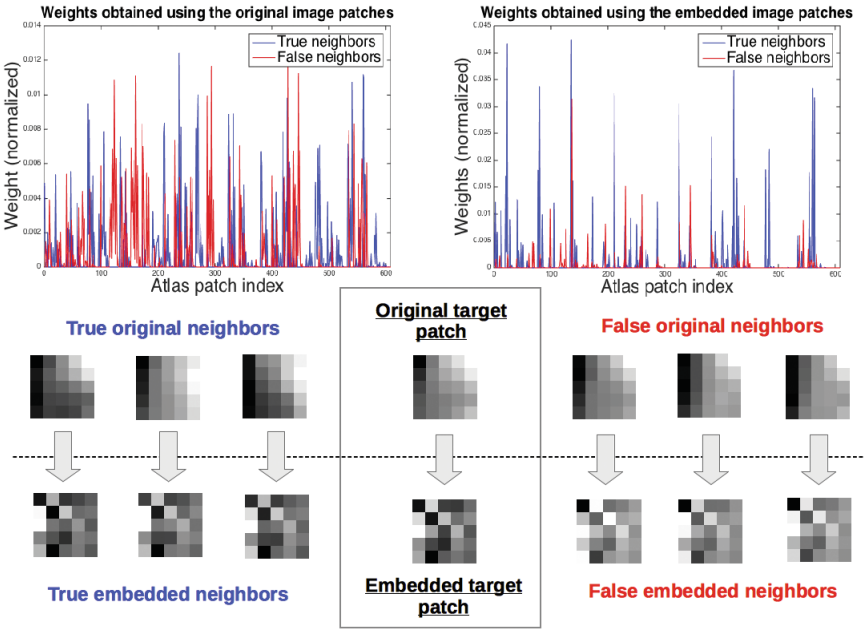


Fig. 1. Weights obtained using the example target patch in the middle box. **Top row:** estimated weights for the neighboring atlas patches using the original patches (left) and the embedded patches (right). Vertical axis represent the weights and horizontal axis represent the atlas patch index. Using the embedded patches, false neighbors (in red) accumulate less weight than true neighbors (in blue). **Middle row:** Original target patch and some true and false neighbors with high weights (note that we only show the center slice of a cubic $5 \times 5 \times 5$ patch). **Bottom row:** Most discriminative 25 coordinates of the embedded patches, arranged in a 5×5 patch (we show the first 25 coordinates for the convenience of displaying them in a 5×5 patch). (Color figure online)

– The proposed method provides a compact representation with a much lower dimensionality than the original patch.

The remainder of the paper is organised as follows: in Sect. 2 we describe the method. In Sect. 3 we present the experiments and results and, finally in Sect. 4 we outline the conclusions.

2 Method

2.1 Patch-Based Label Fusion

Consider a set of n atlas images and label maps, denoted as $\left\{A^i, L^i\right\}_{i=1}^n$, *that have been previously registered to a common space*, denoted as Ω_C (for instance by groupwise non-rigid registration [20]). Therefore, A_x^i denotes the image intensity value at voxel $x \in \Omega_C$ of the i-th atlas, and $L_x^i \in \{0, 1\}$ denotes whether x

belongs (1) or not (0) to the area of interest to be segmented. We denote as T the to-be-segmented target image, after being registered to the common space.

Weighted voting label fusion estimates the label at each target point, denoted as $\hat{F}_x$, as a weighted average of the neighboring atlas labels[1]. That is,

$$\hat{F}_x = \sum_{i=1}^{n} \sum_{y \in \mathcal{N}_x} \omega_y^i L_y^i \tag{1}$$

where ω_y^i denotes the weight of atlas label L_y^i at position $y \in \mathcal{N}_x$ on the i-th atlas, with $\mathcal{N}_x$ denoting the local neighborhood of point x. The local neighborhood $\mathcal{N}_x$ consists of the patches in a cubic neighborhood of a certain radius from point x.

The eventual segmentation performance depends on the ability of the label fusion method to identify the true anatomical neighbors of the to-be-segmented target point among the atlas labels. In particular, PBLF assigns higher weights to the atlas locations with higher local image similarity to the target point [3, 4, 12, 17, 22]. As for the image similarity measures, the Gaussian kernel is widely used to estimate the weights [4, 12]. That is,

$$\omega_y^i = \exp\left(-\|\boldsymbol{t}_x - \boldsymbol{a}_y^i\|^2/\gamma\right), \tag{2}$$

where $\boldsymbol{t}_x, \boldsymbol{a}_y^i \in \mathbb{R}^p$ denote the vectors of the target and the (i-th) atlas image patch centered at x and $y \in \mathcal{N}_x$, respectively, γ is a normalization factor, which is set here as in [4] as $\gamma = \min_{y,i} \|\boldsymbol{t}_x - \boldsymbol{a}_y^i\|^2$, and $\|\cdot\|$ is the Euclidean norm.

2.2 Learning Discriminative Embeddings

PBLF assumes that the higher the similarity between the target patch $\boldsymbol{t}_x$ and an atlas patch $\boldsymbol{a}_y^i$, the higher the likelihood that they share the same label. This simplistic assumption, as expressed in Eq. (2), considers that all the features are equally relevant in capturing the anatomical similarity.

Our goal is to learn a transformation for each point $x \in \Omega_C$, denoted by the matrix $\mathbf{P} \in \mathbb{R}^{p \times d}$, to a lower-dimensional space so that the weights obtained using the transformed patches successfully identify the *anatomically* equivalent patches (rather than the *apparently* similar).

We use all the available atlases as training set, where the image patch from each of the atlases is used as target patch, denoted as a_x^t, and the neighboring patches from the rest of the atlases are used as atlas patches, denoted as a_y^i (with $i \neq t$ and $y \in \mathcal{N}_x$).

The training is performed in the common space. This means that all the training images are registered to a template image (built e.g., by groupwise registration [20]). We learn a different transformation (denoted as $\mathbf{P}$ below) for each point in the common space.

We seek the transformation $\mathbf{P}$ that simultaneously maximizes the distance with the false neighbors and minimizes the distance with the true neighbors,

[1] The estimated label map in the common space $\hat{F}$ is finally transformed back to the original target space and thresholded to obtain the binary labels.

where the true (false) neighbors are the sub-set of positive (negative) samples with higher appearance similarity with the target patch. This can be expressed as follows:

$$\max_{\mathbf{P}} \frac{\sum_{t=1}^{n} \sum_{(i,y)\in\mathcal{F}_x^t} \|\mathbf{P}^\top \boldsymbol{a}_x^t - \mathbf{P}^\top \boldsymbol{a}_y^i\|^2 \, u_y^{t,i}}{\sum_{t=1}^{n} \sum_{(i',y')\in\mathcal{T}_x^t} \|\mathbf{P}^\top \boldsymbol{a}_x^t - \mathbf{P}^\top \boldsymbol{a}_{y'}^{i'}\|^2 \, v_{y'}^{t,i'}}, \tag{3}$$

where $u_y^{t,i}$ and $v_{y'}^{t,i'}$ are the weights identifying the false and true neighbors, respectively (i.e., $u_y^{t,i} > 0$ only for those negative samples with higher appearance similarity to the target patch) and $\mathcal{F}_x^t = \left\{(i,y)\,|L_y^i \neq L_x^t, i \neq t, y \in \mathcal{N}_x\right\}$ is the set of negative samples (the set of positive samples $\mathcal{T}_x^t$ is similarly defined). We compute the weights $u_y^{t,i}$ and $v_{y'}^{t,i'}$ for each target patch in our training set by restricting Eq. (2) to the set of its positive and negative samples, respectively.

The intuition of Eq. (3) is to seek the linear transformation that emphasizes the characteristic differences between false neighbors, so that they are less likely to mislead label fusion, while at the same time downscaling the characteristic differences between true neighbors, so that they end up accumulating more weight. This optimization is somewhat related to linear discriminant analysis (LDA) [5] in the sense that it distinguishes both positive and negative samples for learning the transformation. However, the objective function of LDA is different since it seeks to maximize the between-class scatter and minimize the within-class scatter. In this regard, our approach is more related to manifold learning methods such as locality preserving projections (LPP) [8] and neighborhood preserving embeddings (NPE) [9], but with the difference that we not only minimize the distance with the true neighbors but also jointly maximize the distance with the false neighbors.

Equation (3) can be expressed more compactly as follows:

$$\max_{\mathbf{P}} \frac{Tr\left[\mathbf{P}^\top \mathbf{E}_F \mathbf{U} \mathbf{E}_F^\top \mathbf{P}\right]}{Tr\left[\mathbf{P}^\top \mathbf{E}_T \mathbf{V} \mathbf{E}_T^\top \mathbf{P}\right]}, \tag{4}$$

where $\mathbf{E}_F = \left[\ldots, \boldsymbol{e}_y^{t,i}, \ldots\right] \in \mathbb{R}^{p\times q}$ is a matrix with the columns containing vectors of differences between pairs of false neighbors, i.e., $\boldsymbol{e}_y^{t,i} = \boldsymbol{a}_x^t - \boldsymbol{a}_y^i$ (with $t = 1, \ldots, n$ and $(i,y) \in \mathcal{F}_x^t$), and $\mathbf{U} \in \mathbb{R}^{q\times q}$ is a diagonal matrix with the corresponding weights $u_y^{t,i}$. ($\mathbf{E}_T$ and $\mathbf{V}$ are similarly defined using the differences between true neighbors and their weights, respectively). The larger dimension of the matrices is $q = nk$, where n is the number of atlases in the training set and k is the expected number of false/true neighbors for each target patch. The solution of Eq. (4) can be found according to the following generalized eigenvalue problem:

$$\left(\mathbf{E}_T \mathbf{V} \mathbf{E}_T^\top\right)^{-1} \mathbf{E}_F \mathbf{U} \mathbf{E}_F^\top \boldsymbol{p} = \lambda \boldsymbol{p}, \tag{5}$$

where the desired embedding $\mathbf{P} = [\boldsymbol{p}_1, \ldots, \boldsymbol{p}_d]$ is composed of the $d < p$ eigenvectors with the largest eigenvalues, where d is the desired number of dimensions.

To avoid the possible over-fitting problem, in Eq. (5) we substitute the matrices $\mathbf{S}_F \equiv \mathbf{E}_F \mathbf{U} \mathbf{E}_F^\top$ and $\mathbf{S}_T \equiv \mathbf{E}_T \mathbf{V} \mathbf{E}_T^\top$ by their regularized counterparts [6], as follows:

$$\mathbf{R}_F = (1-\alpha)\,\mathbf{S}_F + \frac{\alpha}{p}\,Tr\left[\mathbf{S}_F\right] I \tag{6}$$

(and similarly for $\mathbf{S}_T$), where $0 \leq \alpha \leq 1$ is a parameter controlling the amount of regularization and I is the identity matrix.

In the testing stage, we estimate the label map $\hat{F}$ of a new target image T according to the following steps:

1. We register the target image to the common space.
2. For each point $x \in \Omega_C$, we extract the surrounding target image patch $\boldsymbol{t}_x$ and the set of neighboring atlas image patches from all the atlases, i.e., $\boldsymbol{a}_y^i$, with $i = 1, \ldots, n$ and $y \in \mathcal{N}_x$, where $\mathcal{N}_x$ is a cubic neighborhood of a certain radius from point x.
3. We estimate the weights ω_y^i for each atlas patch according to Eq. (2) using the embedded target and atlas image patches, $\mathbf{P}^\top \boldsymbol{t}_x$ and $\mathbf{P}^\top \boldsymbol{a}_y^i$, respectively.
4. We estimate the label $\hat{F}_x$ on each target point by fusing the atlas labels according to Eq. (1) using the weights ω_y^i estimated in the previous step.
5. We transform the estimated target label map $\hat{F}$ back to the original target space by using the inverse spatial transformation to the common space.

3 Experiments and Results

We compare our proposed approach to the following methods: (i) majority voting (**MV**) [11], which assigns each target label as the label appearing most frequently among the corresponding atlas labels, and (ii) non-local weighted voting (**NLWV**) [4,12], which uses Eqs. (1) and (2) to estimate the labels and weights, respectively. We apply the proposed discriminative dimensionality reduction on the NLWV pipeline (**DDRNL**), hence we can clearly evaluate the effect of embedding the patches by comparing with the baseline NLWV.

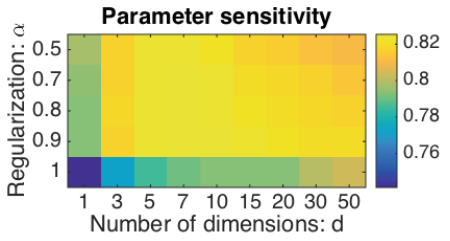
In all the methods, we perform 5-fold cross-validation experiments, where one of the folds is considered as the target images and the rest of the folds as the atlas images. In each fold, the projection matrices $\mathbf{P}$ learned from the atlas images (one projection for each point in the common space) are used to segment the target images. Target images are segmented in the common space and evaluated in the target space by using the Dice similarity coefficient (DSC) with the ground-truth label maps. We use the group-wise non-rigid registration method in [20] to create the template image defining the common space, and diffeomorphic demons [19] to register the target images to the common space. In both NLWV and DDRNL, we use a patch size of $5 \times 5 \times 5$ and a cubic search neighborhood $\mathcal{N}_x$ of $3 \times 3 \times 3$. We evaluate the proposed method on brain MR image segmentation experiments on the ADNI[2] and SATA[3] datasets.

[2] http://www.adni-info.org/.
[3] http://masi.vuse.vanderbilt.edu/workshop2013/index.php/Main_Page.

3.1 ADNI Dataset

The ADNI dataset is provided by the Alzheimer's Disease Neuroimaging Initiative and contains the segmentations of the left and right hippocampi. We use images from 40 randomly selected subjects, where the size of each image is $256 \times 256 \times 256$.

We first conduct a sensitivity analysis on a sub-set of 10 randomly selected images. Figure 2(a) shows the sensitivity to the regularization parameter α and the number of dimensions d. Based on these results we choose the values of the regularization parameter $\alpha = 0.9$ and the number of dimensions $d = 7$ (which is considerably lower than the 125 dimensions of the original $5 \times 5 \times 5$ patches). In Fig. 2(b) we show the average DSC (%) ($\pm$std) in segmenting the left and right hippocampus across the 40 images. As we can see, our proposed method obtains a considerable improvement of $>1.4\,\%$ with respect to the NLWV baseline. Results of MV provide a reference of what can be obtained by the only means of non-rigid registration without using any confidence estimates to weigh the atlases.



(a)

Method	R HC	L HC
MV	76.15 ± 4.04	76.77 ± 3.65
NLWV	79.25 ± 3.25	79.80 ± 3.07
DDRNL	$\mathbf{80.95 \pm 2.66}$	$\mathbf{81.27 \pm 2.25}$

(b)

Fig. 2. (a)Parameter sensitivity analysis and, (b) quantitative segmentation results.

3.2 SATA Dataset

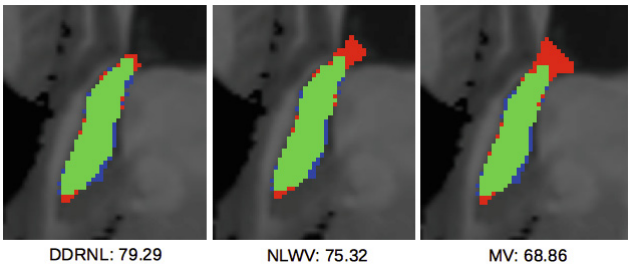
The SATA dataset is composed of 35 images with a resolution of $1 \times 1 \times 1$ mm and contains the segmentation of 16 mid-brain structures. We will focus on the 10 smallest structures since they tend to be more sensitive to registration errors and hence, more challenging to segment. The segmented structures include the right and left parts of: *accumbens, amygdala, pallidum, caudate* and *hippocampus*.

Figure 3(a) shows the average segmentation performance by each method across the 35 images in each structure. Here, we have used the same parameter values $d = 7$ and $\alpha = 0.9$ as in the previous experiments. We have grouped the left and right parts of each structure, so each cell contains the average of 70 segmentations. As we can see, our proposed method obtains a consistent improvement of $\sim 1\,\%$ and even $>1\,\%$ in some structures with respect to the baseline NLWV. Again, MV provides a reference of what can be achieved without resorting to image similarity measurements to weight the atlas contributions.

To gain further insight, in Fig. 3(b) we show the estimated hippocampus segmentations by each method on an example target image. As we can see by the MV results, the head of the hippocampus in this target image is consistently over-segmented by the majority of atlases. NLWV can partially correct this effect by using image-patch similarity to discard some misleading atlases. The proposed method can solve this over-segmentation by using only the discriminative image characteristics in the patch similarity comparisons.

Method	ACC	AMYG	PAL	CN	HC
MV	67.43 ± 10.84	71.35 ± 8.36	71.16 ± 15.43	79.42 ± 8.50	79.06 ± 6.22
NLWV	73.74 ± 6.00	73.76 ± 8.02	82.04 ± 7.15	86.89 ± 3.95	83.09 ± 4.08
DDRNL	$\mathbf{75.61 \pm 4.95}$	$\mathbf{74.59 \pm 8.47}$	$\mathbf{84.30 \pm 4.78}$	$\mathbf{87.65 \pm 3.47}$	$\mathbf{83.85 \pm 3.35}$

(a)



(b)

Fig. 3. (a)Quantitative segmentation results and, (b) an example of qualitative segmentation results on the hippocampus, where green indicates coincidence with the ground-truth labels (i.e., true positive), red indicates excessive segmentation (i.e., false positives) and blue indicates insufficient segmentation (i.e., false negatives). (Color figure online)

4 Discussion and Conclusions

We have presented a dimensionality reduction method to learn optimal patch representations for label fusion that can be plugged into any existing PBLF method. Such representations are learned in the common space so that they can be readily applied to any target image that has been previously aligned to the common space.

Since the proposed method performs label fusion in the common space, the target image needs only to be registered once. This represents a computational advantage with respect to performing it in the target space, since the latter one requires each atlas to be independently registered to the target image space. However, there is some evidence pointing out to the superior performance of using the target space [18]. A possible reason is that pairwise registration accuracy through the common space might not be as accurate as directly warping

the atlas images onto the target image. As a future work, we plan to adapt our method to perform label fusion in the target space.

It is worth noting that the proposed method requires a fair amount of regularization ($\alpha = 0.9$). We believe that this is due to the high complexity involved in learning a different model for each point. A possible solution would be to group the points into perceptually similar regions and learn a single classifier per each region instead of per each point.

We have shown the benefit of the proposed patch representations in the segmentation of several brain structures. We achieve considerable improvements using a compact representation of only 7 dimensions, compared to the 125 dimensions of the original patch.

References

1. Aljabar, P., Heckemann, R.A., Hammers, A., Hajnal, R.V., Rueckert, D.: Multi-atlas based segmentation of brain images: atlas selection and its effect on accuracy. NeuroImage **46**, 726–38 (2009)
2. Amieva, H., Le Goff, M., Millet, X., Orgogozo, J., Prs, K., Barberger-Gateau, P., Jacqmin-Gadda, H., Dartigues, J.: Prodromal Alzheimer's disease: successive emergence of the clinical symptoms. Ann. Neurol. **64**, 492–498 (2008)
3. Artaechevarria, X., Munoz-Barrutia, A., Ortiz-de Solorzano, C.: Combination strategies in multi-atlas image segmentation: application to brain MR data. IEEE Trans. Med. Imaging **28**(8), 1266–1277 (2009)
4. Coupé, P., Manjón, J., Fonov, V., Pruessner, J., Robles, M., Collins, D.L.: Patch-based segmentation using expert priors: application to hippocampus and ventricle segmentation. NeuroImage **54**(2), 940–954 (2011)
5. Duda, R.O., Hart, P.E., Stork, D.H.: Pattern Classification. Wiley Interscience, New York (2000)
6. Friedman, J.H.: Regularized discriminant analysis. J. Am. Stat. Assoc. **84**(405), 165–175 (1989)
7. Goshtasby, A.A.: 2-D and 3-D Image Registration: For Medical, Remote Sensing, and Industrial Applications. Wiley-Interscience, Hoboken (2005)
8. He, X.F., Niyogi, P.: Locality preserving projections. In: NIPS (2003)
9. He, X., Cai, D., Yan, S., Zhang, H.: Neighborhood preserving embedding. In: ICCV (2005)
10. Heckemann, R.A., Hajnal, J.V., Aljabar, P., Rueckert, D., Hammers, A.: Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. NeuroImage **33**(1), 115–126 (2006)
11. Rohlfing, T., Brandt, R., Menzel, R., Maurer Jr., C.R.: Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. NeuroImage **21**(4), 1428–1442 (2004)
12. Rousseau, F., Habas, P.A., Studholme, C.: A supervised patch-based approach for human brain labeling. IEEE Trans. Med. Imaging **30**(10), 1852–1862 (2011)
13. Sanroma, G., Wu, G., Gao, Y., Shen, D.: Learning-based atlas selection for multiple-atlas segmentation. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 3111–3117. IEEE (2014)
14. Sanroma, G., Wu, G., Gao, Y., Shen, D.: Learning to rank atlases for multiple-atlas segmentation. IEEE Trans. Med. Imaging **33**(10), 1939–1953 (2014)

15. Sanroma, G., Wu, G., Gao, Y., Thung, K.H., Guo, Y., Shen, D.: A transversal approach for patch-based label fusion via matrix completion. Med. Image Anal. **24**(1), 135–148 (2015)
16. Sanroma, G., Wu, G., Thung, K., Guo, Y., Shen, D.: Novel multi-atlas segmentation by matrix completion. In: Wu, G., Zhang, D., Zhou, L. (eds.) MLMI 2014. LNCS, vol. 8679, pp. 207–214. Springer, Heidelberg (2014)
17. Tong, T., Wolz, R., Hajnal, J.V., Rueckert, D.: Segmentation of brain MR images via sparse patch representation. In: STMI (2012)
18. Tong, T., Wolz, R., Coupé, P., Hajnal, J.V., Rueckert, D.: Segmentation of MR images via discriminative dictionary learning and sparse coding: application to hippocampus labeling. NeuroImage **76**, 11–23 (2013)
19. Vercauteren, T., Pennec, X., Perchant, A., Ayache, N.: Diffeomorphic demons: efficient non-parametric image registration. NeuroImage **45**(1, Supplement 1), S61–S72 (2009)
20. Wu, G., Jia, H., Wang, Q., Shen, D.: Sharp mean: groupwise registration guided by sharp mean image and tree-based registration. NeuroImage **56**(4), 1968–1981 (2012)
21. Wu, G., Kim, M., Sanroma, G., Wang, Q., Munsell, B.C., Shen, D., Initiative, A.D.N., et al.: Hierarchical multi-atlas label fusion with multi-scale feature representation and label-specific patch partition. NeuroImage **106**, 34–46 (2015)
22. Zhang, D., Guo, Q., Wu, G., Shen, D.: Sparse patch-based label fusion for multi-atlas segmentation. In: Yap, P.-T., Liu, T., Shen, D., Westin, C.-F., Shen, L. (eds.) MBIA 2012. LNCS, vol. 7509, pp. 94–102. Springer, Heidelberg (2012)

Author Index

Achterberg, Hakim C. 85
Aksan, Emre 25
Alansary, Amir 13
Alexander, Daniel C. 35
Aljundi, Rahaf 74
Ashburner, John 45

Benkarim, Oualid M. 94

Cardoso, M. Jorge 45

de Bruijne, Marleen 85
Durand-Dubief, Françoise 57

Firat, Orhan 25

Glocker, Ben 13
González Ballester, Miguel Ángel 94

Hajnal, Joseph V. 13
Hannoun, Salem 57

Kadoury, Samuel 65
Kainz, Bernhard 13
Keraudren, Kevin 13
Kocevar, Gabriel 57

Lartizien, Carole 74
Lee, Matthew 13
Lehaire, Jérôme 74
Loktyushin, Alexander 3
Lombaert, Hervé 65
Lorenzi, Marco 35

Malamateniou, Christina 13
Mendelson, Alex 45
Modat, Marc 45

Ourselin, Sebastien 35, 45
Oztekin, Ilke 25

Piella, Gemma 94
Prost-Boucle, Fabrice 74

Rouvière, Olivier 74
Rueckert, Daniel 13
Rutherford, Mary 13

Sanroma, Gerard 94
Sappey-Marinier, Dominique 57
Scheffler, Klaus 3
Schölkopf, Bernhard 3
Schuler, Christian 3
Shakeri, Mahsa 65
Shen, Dinggang 94
Stamile, Claudio 57

van Opbroek, Annegreet 85

Wu, Guorong 94

Yarman Vural, Fatos T. 25
Young, Jonathan 45

Zhu, Xiaofeng 94
Ziegler, Gabriel 35