Automated Extraction of Biomarkers for Alzheimer's Disease from Brain Magnetic Resonance Images

Robin Wolz

A dissertation submitted in partial fulfilment
of the requirements for the degree of
Doctor of Philosophy
of
Imperial College London

February 2011
Department of Computing
Imperial College London

To my mother

Declaration of originality

I hereby declare that the work described in this thesis is my own, except where specifically acknowledged.

Robin Wolz

Abstract

In this work, different techniques for the automated extraction of biomarkers for Alzheimer's disease (AD) from brain magnetic resonance imaging (MRI) are proposed. The described work forms part of PredictAD (www.predictad.eu), a joined European research project aiming at the identification of a unified biomarker for AD combining different clinical and imaging measurements. Two different approaches are followed in this thesis towards the extraction of MRI-based biomarkers: (I) the extraction of traditional morphological biomarkers based on neuronatomical structures and (II) the extraction of data-driven biomarkers applying machine-learning techniques. A novel method for a unified and automated estimation of structural volumes and volume changes is proposed. Furthermore, a new technique that allows the low-dimensional representation of a high-dimensional image population for data analysis and visualization is described. All presented methods are evaluated on images from the Alzheimer's Disease Neuroimaging Initiative (ADNI), providing a large and diverse clinical database. A rigorous evaluation of the power of all identified biomarkers to discriminate between clinical subject groups is presented. In addition, the agreement of automatically derived volumes with reference labels as well as the power of the proposed method to measure changes in a subject's atrophy rate are assessed. The proposed methods compare favorably to state-of-the art techniques in neuroimaging in terms of accuracy, robustness and run-time.

Acknowledgments

I would first of all like to thank my supervisor Daniel Rueckert for his help and support throughout this thesis. His inspiration and constant motivation made sure that working on this project always was both a scientific challenge and a pleasure. I am also grateful to Paul Aljabar for many fruitful discussions and his patience with all my questions.

I want to thank the whole PredictAD team and in particular Jyrki Lötjönen for providing a motivating research environment. Furthermore, I am thankful to the members of the Imperial Imaging Meeting, especially Jo Hajnal, Alex Hammers and Rolf Heckemann.

Finally, I want to thank the whole BioMedIA group for a great atmosphere as well as my friends and family for constant motivation.

Contents

1 Introduction **1**
- 1.1 Biomarkers for AD . . . 2
- 1.2 PredictAD . . . 3
- 1.3 Imaging data . . . 3
- 1.4 Thesis contributions . . . 4
  - 1.4.1 Multi-atlas segmentation of diverse populations with automated intensity-refinement . . . 5
  - 1.4.2 Consistent segmentation of image sequences to measure atrophy 5
  - 1.4.3 Manifold learning combining imaging with non-imaging information to classify subjects . . . 6
  - 1.4.4 Comprehensive analysis of the developed biomarkers . . . 6

2 Background **7**
- 2.1 Brain atlases . . . 7
  - 2.1.1 Hammers brain atlases . . . 8
- 2.2 Atlas based brain segmentation . . . 8
  - 2.2.1 Atlas segmentation incorporating intensity modeling . . . 9
  - 2.2.2 Multi-atlas segmentation . . . 11
  - 2.2.3 Multi-atlas segmentation incorporating intensity modeling . . . 13
- 2.3 Atrophy measurement . . . 13
- 2.4 Manifold learning . . . 15
  - 2.4.1 Dense spectral techniques . . . 16
  - 2.4.2 Sparse spectral techniques . . . 18
  - 2.4.3 Application of manifold learning . . . 20

3 Automated intensity-refinement with multi-atlas label propagation **23**
- 3.1 Introduction . . . 24
- 3.2 Method . . . 25
  - 3.2.1 Estimation of a subject-specific data term . . . 26
  - 3.2.2 Smoothness term . . . 29
- 3.3 Data and Results . . . 29
  - 3.3.1 Comparison with manually labeled data . . . 30
  - 3.3.2 Visual inspection . . . 31
- 3.4 Conclusion . . . 32

4 LEAP: Learning Embeddings for Atlas Propagation **34**
- 4.1 Introduction . . . 35
- 4.2 Materials and Methods . . . 38

4.2.1 Subjects . . . 38
4.2.2 Atlases . . . 39
4.2.3 Overview of the method . . . 39
4.2.4 Graph Construction and Manifold Embedding . . . 40
4.2.5 Segmentation Propagation in the Learned Manifold . . . 42
4.2.6 Multi-atlas propagation and segmentation refinement . . . 43
4.3 Experiments and Results . . . 44
4.3.1 Image similarities . . . 44
4.3.2 Coordinate system embedding . . . 45
4.3.3 Evaluation of hippocampus segmentations . . . 46
4.3.4 Volume measurements . . . 50
4.3.5 Segmentation of 83 brain structures . . . 52
4.4 Discussion and Conclusion . . . 58

5 Consistent segmentation of longitudinal images to measure atrophy 61
5.1 Introduction . . . 62
5.2 Materials and Methods . . . 64
5.2.1 Image data . . . 64
5.2.2 Hippocampus atlases . . . 65
5.2.3 4D image segmentation with graph-cuts . . . 66
5.3 Experiments and Results . . . 68
5.3.1 Hippocampal atrophy after 12 and 24 months . . . 69
5.3.2 Correlation with clinical values . . . 71
5.3.3 ApoE genotype . . . 72
5.3.4 Discrimination between clinical groups based on atrophy . . . 73
5.3.5 Sample size calculation . . . 74
5.3.6 Segmentation accuracy . . . 75
5.4 Discussion and Conclusion . . . 78

6 Biomarker extraction from manifold learning 82
6.1 Introduction . . . 83
6.2 Method . . . 85
6.2.1 Manifold learning using pairwise image similarities . . . 85
6.2.2 Manifold learning incorporating non-imaging information . . . 87
6.2.3 Extraction of biomarkers . . . 91
6.3 Data and Results . . . 92
6.3.1 Subjects . . . 92
6.3.2 Pairwise image similarities . . . 93
6.3.3 Experiments . . . 94
6.3.4 Parameter settings . . . 96
6.3.5 Classification . . . 97
6.3.6 Regression . . . 98
6.3.7 Alternative approaches to incorporate metadata . . . 98
6.4 Discussion . . . 101

7 Manifold learning incorporating longitudinal data **106**
7.1 Introduction . . . 107
7.2 Method . . . 107
7.2.1 Manifold learning for cross-sectional data . . . 107
7.2.2 Manifold learning for longitudinal data . . . 108
7.3 Experiments and results . . . 109
7.3.1 Subjects . . . 109
7.3.2 Parameter settings . . . 109
7.3.3 Classification . . . 110
7.4 Discussion and conclusion . . . 113

8 Comprehensive analysis of MR-derived biomarkers **115**
8.1 Introduction . . . 116
8.2 Materials and Methods . . . 117
8.2.1 Subjects . . . 117
8.2.2 Statistical analysis . . . 120
8.2.3 Classification . . . 121
8.3 Experiments and results . . . 122
8.3.1 Image sets . . . 122
8.3.2 Classification results using dataset I . . . 122
8.3.3 Classification results using dataset II . . . 124
8.4 Discussion . . . 125

9 Summary and Conclusion **130**
9.1 Classification performance . . . 131
9.2 Performance based on other measures . . . 133
9.2.1 Label overlaps . . . 134
9.2.2 Sample size . . . 135
9.3 Conclusion . . . 137
9.4 Future work . . . 138

10 Publications **140**
10.1 Book Chapter . . . 140
10.2 Journal Publications . . . 140
10.3 Conference Proceedings . . . 141
10.4 Conference Abstracts . . . 143
10.5 Patent . . . 144

A ADNI **145**
A.1 MR image acquisition . . . 145
A.1.1 Hippocampus reference labels . . . 146

B Hammers atlas **147**

List of Tables

1.1 1.5T ADNI MRI images available in January 2011. The bottom part shows the visits when cognitive tests (MMSE, CDG) and the CSF-based markers are taken for CN/MCI/AD. x: measure available, -: no measure available. 4

3.1 Average SI overlap for hippocampus segmentation. 30

4.1 Information relating to the subjects whose images were used in this study. 38
4.2 Characteristics of the subjects used for comparison between manual and automatic segmentation 47
4.3 Similarity index (SI) for hippocampus segmentation. 48
4.4 Classification accuracy (ACC), sensitivity (SEN) and specificity (SPE) achieved with support vector machines (SVMs) based on automatically determined volumes. Results are displayed for all structures for which the classification accuracy of at least one comparison performed better than chance. 54
4.5 Classification results achieved with the combination of different structural volumes. Support vector machines (SVM) and AdaBoost were used on different sets of volumes to perform classification. 56
4.6 Classification accuracies achieved for selected structural volumes. The top part of the table shows results after direct propagation of atlas images, the bottom part shows the results after applying LEAP where pairwise similarity is measured in a region around the structure of interest. 58

5.1 Clinical and demographical overview of the study population. Mean age of 75.3 ± 6.6 years and mean time between both scans of 12.96 ± 1.32 months for the whole population does not vary between subject groups. 65
5.2 Subpopulation for which three timepoints were available. The number of subjects, number of females and average change in MMSE and CDR during 24 months are given for the six subject groups. 65
5.3 Hippocampal atrophy rates (%) in 555 subjects over 12 months. Number of subjects are given in parentheses. Mean±std 69
5.4 Hippocampal atrophy rates (%) in 352 subjects over 24 months. Number of subjects are given in parentheses. Mean±std 70
5.5 Correlation of 12-month atrophy rates with clinical values. Number of subjects are given in parentheses. (a: $p<0.001$, b: $p<0.01$) 72

5.6 T-statistics for the hypothesis of atrophy rates over 12 months in ε3/3 and ε3/4 carriers come from the same distribution. The number of subjects carrying E3 and E4 respectively is given in parentheses. a: p<0.001 72
5.7 Classification results using automatically determined atrophy rates after 12 months and after 24 months in parentheses. 74
5.8 Estimated sample sizes for both arms that would be needed to detect a 25% reduction in atrophy in the AD and MCI groups in intervals of 12 and 24 months. 75
5.9 Average atrophy rates (%) for the subset of image Set 1 for which hippocampal label maps were provided by ADNI. Atrophy rates based on these label maps are compared to automatically determined rates based on the proposed method. Numbers of subjects are given in parentheses. mean±std 76

6.1 Subject data of the study subjects are shown for the different groups. Non-imaging metadata in the form of ApoE genotype and Aβ_{42} concentration as well as the derived imaging metadata, hippocampus volume, are presented. Carriers of the ApoE ε2/ε4 alleles are shown. The remaining subjects only carry the ε3 allele. There is no significant difference in age between the clinical groups with an average age of 74.95±7.03 years. 93
6.2 Correct classification accuracy (ACC), sensitivity (SEN) and specificity (SPE) (%) for classic Laplacian eigenmaps (LE, I) and the extended version E-LE incorporating different types of meta-information (II-VI). P-values for the difference between methods I-VI and method I (p_1) and method VI (p_2) are presented. † stands for p<0.001 The results for method V are significantly different from all other results with p<0.001 apart from method VI. The bottom rows of the table present classification rates when using different types of metadata only. 99
6.3 Statistics from regressing MMSE versus d manifold coordinates using a multiple linear model. An improvement of statistics can be observed when incorporating metadata into the manifold learning process. Results are presented for d=15/d=1 100
6.4 Classification accuracy (ACC), sensitivity (SEN) and specificity (SPE) when incorporating metadata for classification into the SVM featurevector. p<0.001 for differences with the according methods (III, IV) in Table 6.2 are labeled with a **bold** classification accuracy. 100

7.1 Correct classification results in percentages using different feature vectors based on scans' coordinates in the learned manifolds (Section 7.2.2). Vector A is based on baseline features only. For vector B, baseline and follow-up scans (after 12 or 24 months) are together embedded in one manifold. Vector C consists of features taken from the baseline embedding and a separate embedding of longitudinal image differences. Average classification rates (ACC), sensitivity (SEN) and specificity (SPE) are displayed when varying the dimension of the manifold $l \in [6, 15]$. . . 110

7.2 Classification results based on hippocampal baseline volume (Chapter 4) and atrophy (Chapter 5) over 12 and 24 months. The second part of the table shows the correlation of coordinates in the learned manifolds with baseline volume and atrophy. $d = 20$ coordinates of $\mathbf{y}_{i0}^{X_0}$, $\mathbf{y}_{i\Delta J}^{X_{\Delta J}}$ and $\mathbf{y}_{iJ}^{X_0 \cup X_j}$ are used to determine r. a: $p < 10^{-4}$ 112

8.1 Demographic and clinical data of the study subjects. Level of significance is set to p < 0.05. *Different between the groups. [1]Different from controls. [2]Different from all other groups. 122
8.2 CN vs AD. Accuracy (ACC), sensitivity (SEN) and specificity (SPE) are presented for hippocampal volume (HC), hippocampal atrophy (HA), manifold learning (MBL), tensor-based morphometry and the combination of all features. Features are available at baseline (BL), month 12 (m12) and month 24 (m24). 123
8.3 S-MCI vs P-MCI. Accuracy (ACC), sensitivity (SEN) and specificity (SPE) are presented for hippocampal volume (HC), hippocampal atrophy (HA), manifold learning (MBL), tensor-based morphometry and the combination of all features. Features are available at baseline (BL), month 12 (m12) and month 24 (m24). 124
8.4 CN vs P-MCI. Accuracy (ACC), sensitivity (SEN) and specificity (SPE) are presented for hippocampal volume (HC), hippocampal atrophy (HA), manifold learning (MBL), tensor-based morphometry and the combination of all features. Features are available at baseline (BL), month 12 (m12) and month 24 (m24). 125
8.5 CN vs AD. Results for the combination of different feature sets are presented. The used feature sets include cortical thickness (CTH), hippocampal volume (HC), manifold learning (MBL) and tensor-based morphometry (TBM). 126
8.6 S-MCI vs P-MCI. Results for the combination of different feature sets are presented. The used feature sets include cortical thickness (CTH), hippocampal volume (HC), manifold learning (MBL) and tensor-based morphometry (TBM). 126
8.7 P-MCI vs CN. Results for the combination of different feature sets are presented. The used feature sets include cortical thickness (CTH), hippocampal volume (HC), manifold learning (MBL) and tensor-based morphometry (TBM). 127

9.1 Comparison of classification results achieved with the proposed method to state-of-the art methods. I: AD vs CN, II: P-MCI vs S-MCI, III: P-MCI vs CN. SEN/SPE. Classification accuracy is reported where no SEN/SPE was provided. 132
9.2 Hippocampus label overlap 134
9.3 Atrophy rates over 12 months and resulting sample sizes required to detect a 25% change in atrophy rate with 80% power and 5% significance. All results are based on ADNI data, sample sizes are reported for both arms (AD/MCI). Sample sizes are presented for measured atrophy rates (SS I) and atrophy rates corrected for healthy ageing (SS II). N: Number of subjects used. 136

B.1 83 Structures 148

List of Figures

2.1 Manually delineated structures on a brain atlas 8
2.2 Atlas based brain segmentation: an atlas image is registered with the unseen image and atlas labels are propagated. 9
2.3 Multi-atlas brain segmentation. Multiple atlas images are registered with the unseen images and all atlas labels are propagated to the target. The final segmentation is obtained from fusing the individual segmentations. 12
2.4 In the schematic illustration above, the images $\mathbf{x}_i$, $1 \leq i \leq N$ (left) are compared in pairs and measures of similarity or distance between them are obtained. The measures define a $N \times N$ matrix representing the edge weights in a graph representation of the data. The graph/matrix representation may be either full (dense, $\mathbf{W}$ above) or sparse ($\mathbf{W}'$), illustrations of both cases are shown above. Typically, the eigenvalue-eigenvector structure of the matrix (or of a matrix derived from it) is used to derive a coordinate representation for an embedded manifold representation $\mathbf{y}_i$ of the original data. The first two dimensions of $\mathbf{y}_i$ are schematically shown above. 16
2.5 2D embeddings obtained with different dimensionality reduction techniques to 167 images acquired from AD patients (blue) and 231 images from healthy controls (red). 21

3.1 Segmentation with graph cuts. A graph is defined on the target image in which every voxel is represented by a node. Source and sink nodes represent foreground and background and weights from every node to source and sink are defined according to the energy model. Edges connecting neighboring nodes enforce smoothness. The final segmentation is obtained by finding the minimum cut of the defined graph. 26
3.2 Difference between multi-atlas segmentation and the proposed method for the hippocampus segmentation in 60 test cases. 31
3.3 (a) shows a 3D-rendering of the segmentation result for the proposed method for all structures: thalamus (blue), putamen (yellow), caudate (pink), hippocampus (green), amygdala (red) and nucleus accumbens (turquoise). (b): Transverse section showing segmentation outlines superimposed on an MR image. 31
3.4 (a) shows the segmentation results for multi-atlas segmentation (b): results for the proposed method. (c-d): Subject specific probabilistic atlases for hippocampus and amygdala (a higher intensity encodes a higher probability). 32

4.1 Manually delineated structures on a brain atlas 39
4.2 Process of atlas propagation with LEAP. All labeled (atlases) and unlabeled images are embedded into a low-dimensional manifold (1). The N closest unlabeled images to the labeled images are selected for segmentation (2). The M closest labeled images are registered to each of the selected images (an example for one image is shown in (3)). Intensity refinement is used to obtain label maps for each of the selected images (4). Steps (2) - (4) are iterated until all images are labeled. 40
4.3 The discrimination ability for different chosen feature dimensions among the four subject groups (healthy young, elderly controls, MCI, AD). The best discrimination was achieved using a two dimensional embedding space which therefore was used to define the distances between images. 42
4.4 The MNI152 brain atlas showing the region of interest around the hippocampus that was used for the evaluation of pairwise image similarities 45
4.5 Abscissa and ordinate show first and second coordinates respectively of a low-dimensional embedding space. Embedded are 30 atlases based on healthy subjects and 796 images from elderly dementia patients and age matched control subjects. Details of images showing the hippocampus in example subjects. 46
4.6 Comparison of segmentation results for the right hippocampus on a transverse slice. 47
4.7 Development of segmentation accuracy with increasing distance from the original set of atlases. Each subset of images used for evaluation is represented by one bar plot. 49
4.8 Average hippocampal volumes for manual and automatic segmentation using method IV. 50
4.9 A Bland-Altman plot showing the agreement between volume measurement based on manual- and automatic segmentation of the hippocampus (method IV). The solid line represents the mean and the dashed lines represent ± 1.96 standard deviations. 51
4.10 First two embedding coordinate for 796 ADNI images together with 30 manually labeled atlas images obtained from applying a spectral embedding step to pairwise similarities evaluated over the whole brain after an affine normalization to a template space. 53
4.11 Classification accuracy for three different clinical groupings achieved from 83 delineated brain structures. 55
4.12 Label overlaps (SI) for automated hippocampus segmentation with semi-automated reference segmentations. Compared are segmentations obtained with LEAP when using a similarity measure over the whole brain (blue) to a similarity measure defined in a region around the hippocampus (grey). Results for both approaches are represented for 10 groups of subjects as described before for Figure 4.7. 57

5.1 4D graph cut segmentation: images acquired at two timepoints are connected by additional smoothness constraints (black edges) when compared to a 3D graph cut model. 68
5.2 Segmentation of the right hippocampus in an AD subject. baseline (a) and month 12 follow-up (b) segmentation using 4D graph cuts. 69

5.3 Hippocampal volume loss in % from baseline after 12 and 24 months. Box-and whisker plots for AD, P-MCI, S-MCI, CN. 70
5.4 Comparison of volume loss after 12 months when segmenting two (method a) or three (method b) timepoints simultaneously. Dashed lines represent the 95% confidence interval of the mean (solid line). 71
5.5 ROC curves show the discrimination between subject groups. The area under the curve (AUC) for Controls vs AD, Controls vs MCI, Controls vs P-MCI and P-MCI vs S-MCI are 0.88 (0.92), 0.71 (0.77), 0.83 (0.86), and 0.72 (0.71), respectively. AUC's for rates after 24 months are given in parentheses. 73
5.6 ROC curves show the discrimination between subject groups. The area under the curve AUC for (ADNI labels/4D graph-cuts) Controls vs AD, Controls vs MCI, Controls vs P-MCI and P-MCI vs S-MCI are 0.76/0.87, 0.60/0.70, 0.63/0.77, and 0.58/0.66, respectively. 77

6.1 Weights defined between image nodes $\mathbf{x}_i$ and support nodes representing metadata Z. In the discrete setting (left), equally weighted edges are defined according to Equation 6.5. In the continuous setting (right), weights to both additional nodes are defined according to Equations 6.6 and 6.7. A higher weight is illustrated by a thicker edge. 89
6.2 First two embedding coordinates with varying influence of γ. A high weight leads to an embedding similar to the one obtained with classic Laplacian eigenmaps (c). A very low weight embeds the images mainly based on metadata (a). 90
6.3 Orthogonal views of MNI152 space showing the ROI around hippocampus and amygdala used to evaluate pairwise image similarities. 94
6.4 Standard embedding using Laplacian eigenmaps based on pairwise image similarities only (top). Extended embedding using the proposed method with hippocampal volume as metadata (bottom). 103 AD patients are represented by squares, 116 healthy controls by circles. Hippocampal volume (cm^3) is encoded in the marker color. A SVM separating hyperplane in 2 dimensions is displayed. Misclassified subjects with both approaches are highlighted by a black outline (42 with LE, 31 with E-LE). An improved separating ability can be observed in the extended embedding especially for subjects close to the separating plane in the original embedding. 104
6.5 (a): Classification rate in the training data set when varying the dimension of the low-dimensional manifold between 1 and 50. Results are presented as mean rates over ten bins covering five dimensions each. The high standard deviation observed in the first bin results from a very low classification rate with d=1. (b): Classification rate when varying γ between 1 and 50 evaluated for $d \in [6, 15]$. 105

6.6 Classification accuracy obtained from defining a combined similarity measure incorporating both imaging and non-imaging information before performing manifold learning. AD vs CN: blue; S-MCI vs P-MCI: green; CN vs P-MCI: red. Results with hippocampal volume and Aβ_{42} are presented over an increasing influence of the metadata. The dotted lines indicate the classification accuracy obtained with image similarities only. 105

7.1 2D visualizations of manifolds incorporating longitudinal information. Exemplar images are labeled $\mathbf{x_{ij}}$ and $\mathbf{x_{i\Delta j}}$ with $i = 1, ..., 6$ and $j = 0, 2$ where i represents the subject id and j the visit number. 111
7.2 Box-and-whisker plots for the distance between longitudinal images in a manifold learned from images at several timepoints. 112

8.1 Inclusion / exclusion criteria 118

Chapter 1

Introduction

Alzheimer's disease (AD) is the most common cause of dementia. It is a devastating disease for those who are affected and presents a major burden to caretakers and society. The worldwide prevalence of AD is predicted to quadruple from 26.6 million in 2006 to more than 100 million by the year 2050. Even a modest delay of 1 year in the disease onset and progression could reduce the number of cases by 9 million [20] which makes an early diagnosis paramount.

Genetic risk factors for AD have been identified [92, 71]. A definitive diagnosis, however, requires histological examination of brain tissue. In order to decide on a potential treatment of individuals, the identification of people at risk at an early stage of disease development is required. Mild cognitive impairment (MCI) is a heterogeneous syndrome that increases the risk of developing AD markedly. However, not all MCI subjects convert to AD. A focus in the search for biomarkers of AD type pathology therefore lies in predictors of disease progression among the MCI subjects.

In this thesis, methods for an automated extraction of such biomarkers from brain magnetic resonance imaging (MRI) data are developed. Section 1.1 starts with an overview on different AD-biomarkers proposed. The research presented in this thesis was done as part of the European research project PredictAD which aims at defining a unified biomarker for AD. A brief presentation of the aims of the project are given in Section 1.2. Most analysis presented in this work is based on imaging data from

the Alzheimer's Disease Neuroimaging Initiative (ADNI) which is presented in Section 1.3. The contribution of this thesis is summarized in Section 1.4.

1.1 Biomarkers for AD

The methods used to assess the possibility of a given individual to be affected by dementia can be broadly divided into two categories: (I) psychological tests and (II) quantitative measurements. Psychological tests like the Mini-mental State Examination (MMSE) [54] or the Clinical Dementia Rating (CDR) [110] are used in most memory clinics to assess the cognitive state of a new patient. They typically involve several questions testing the short-term memory of the patient. While an existing impairment can be identified in most cases, a much earlier identification of people at risk is necessary to enable a successful treatment. AD is caused by neurofibrillary tangles and neuritic plaques [19]. Degenerative changes in the human neurotransmitter system lead to atrophy in selected brain regions [149].

A promising approach for detecting the disease at its earliest stage is to study the generation of tangles and plaques. The concentration of the tau-protein and the amyloid-beta-peptide Aβ_{42} in the cerebrospinal fluid (CSF) are commonly associated with the risk of developing AD [138]. While obtaining a CSF sample is invasive, this biomarker can give a good assessment of a patient's state.

A decrease in brain metabolism of glucose and oxygen caused by AD can be identified by Positron Emission Tomography (PET) with the use of a Fludeoxyglucose ^{18}F (FDG) tracer [25]. PET in combination with the Pittsburgh Compound B (PiB) tracer has found recent attention as a biomarker for AD [78]. It selectively binds to Aβ deposits and thereby images beta-amyloid deposits.

Structural images acquired with MRI on the other hand allow to analyze the current state of brain degeneration. The volume of brain structures and their change over time are widely accepted as biomarkers for AD, e.g., [81]. A more detailed introduction to biomarkers for AD can be found in, e.g., [138].

1.2 PredictAD

The work presented in this thesis has been developed during the research project PredictAD (www.predictad.eu). PredictAD is a multinational project funded by the European Union aiming at developing a standardized and objective solution that enables an earlier diagnosis of Alzheimers disease, improved monitoring of treatment efficacy and enhanced cost-effectiveness of diagnostic protocols. Apart from MRI-based biomarkers as discussed in this work, it involves PiB PET, electrophysiological data (TMS/EEG), molecular data, demographic data and clinical tests. The aim is to combine the different biomarkers into a Computer Aided Diagnosis (CAD) tool to assist in clinical decision making [105].

1.3 Imaging data

The evaluation of the methods presented in this work is performed on images obtained from the Alzheimer's Disease Neuroimaging Initiative (ADNI) database [111]. The aim of ADNI is to develop biomarkers of AD in elderly subjects. The primary goal has been to test whether serial MRI, PET, other biological markers, and clinical and neuropsychological assessment can be combined to measure the progression of MCI and AD.

A lumbar puncture is performed in a subset of subjects to extract CSF concentration and ratio data for the following biomarkers: Tau, Aβ_{42}, and P-tau$_{181}$. A full genetic study is employed at baseline, extracting more than 620,000 markers including APOE-genotype which has been associated with AD [92]. The cognitive assessment performed in ADNI includes the widely used MMSE and CDR tests. The bottom part of Table 1.3 gives an overview on measurements taken for the different subject groups at different visits.

In ADNI, MRI scans are taken from all participants in regular intervals. Approximately 200 cognitively normal older individuals are followed for 3 years, 400 people with MCI are followed for 3 years, and 200 people with early AD are followed for 2

Type	Baseline	Month 6	Month 12	Month 18	Month 24	Month 36
Normal	231	215	202	n.a.	174	136
S-MCI	241	198	183	149	131	92
P-MCI	168	160	154	135	121	81
AD	198	166	145	n.a.	111	n.a.
Total	838	739	684	284	537	309
MMSE/CDG	x/x/x	x/x/x	x/x/x	-/x/-	x/x/x	x/x/-
CSF	x/x/x	-/-/-	x/x/x	-/-/-	x/x/x	x/x/-

Table 1.1: 1.5T ADNI MRI images available in January 2011. The bottom part shows the visits when cognitive tests (MMSE, CDG) and the CSF-based markers are taken for CN/MCI/AD. x: measure available, -: no measure available.

years (www.adni-info.org). The clinical group of all subjects is re-assessed at every visit. Retrospectively discriminating between MCI subjects with a stable diagnosis (S-MCI) and progressive MCI (P-MCI) subjects that convert to AD, allows to test the ability of biomarkers measured at baseline to predict such a conversion.

An overview of the 1.5T MR images that were available in January 2011 is presented in Table 1.3. Every chapter of this thesis gives an overview on the particular subset of images used in that analysis. A more detailed description of image acquisition and preprocessing in ADNI is given in Appendix A.1

For a subset of ADNI images, a reference hippocampus segmentation is available which is used for the evaluation of parts of this work. This reference is based on a semi-automatically generated and manually corrected segmentation. A detailed description of the protocol is given in Appendix A.1.1.

1.4 Thesis contributions

This thesis presents methods for an automated extraction of biomarkers from serial MRI images. The types of methods presented can be divided into two categories. The first category includes methods that extract traditional biomarkers based on an automated segmentation of brain structures and their volumes or volume changes. Developed methods that deal with structural brain segmentation are covered in Chapters 3, 4 and 5. The second category of methods employs methods from machine learning

to derive more data-driven biomarkers. Developed methods that use such approaches are presented in Chapters 6 and 7. Chapter 8 presents a comprehensive analysis where the biomarkers developed in this work are combined with other automatically derived MR-biomarkers to test the power of a combined biomarker to classify between different subject groups.

The reminder of this section gives a more detailed overview on the contributions of this thesis.

1.4.1 Multi-atlas segmentation of diverse populations with automated intensity-refinement

Chapter 3 describes a fully automated method to combine multi-atlas label propagation with an intensity-based refinement step based on graph cuts. Building on this, Chapter 4 describes a novel framework to automatically propagate a set of labeled atlases through to a diverse set of images. The presented method can significantly improve segmentation with multi-atlas segmentation in cases where available atlases are based on only a sub-population of the target dataset. It is robust to differences in the MR sequence of images used, and only requires minimal parameter setting. Since the manual labeling of atlas images is time-consuming and expensive, such a framework can be particularly useful in the automated analysis of large diverse clinical image databases as required in, e.g., clinical trials.

1.4.2 Consistent segmentation of image sequences to measure atrophy

Extending on the multi-atlas framework described above, Chapter 5 describes a method for the segmentation of longitudinal image sequences. Measuring longitudinal brain development may allow to draw more accurate conclusions on a subject's clinical state than a cross-sectional comparison alone. For the accurate measurement of volume changes, a consistent segmentation at baseline and follow-up is required. The ap-

proach presented in this work is based on the simultaneous segmentation of all time points in a unified optimization step. The resulting segmentation allows the accurate measurement of atrophy allowing a promising classification accuracy and a high statistical power to reliably measure changes in atrophy rate, a widely used measure of drug efficacy.

1.4.3 Manifold learning combining imaging with non-imaging information to classify subjects

A data-driven approach for the extraction of biomarkers is proposed in Chapters 6 and 7. Manifold learning is applied to a set of brain images, defining a low-dimensional representation of the population. Traditionally based on pairwise similarities between all images, Chapter 6 describes an extension to an established manifold learning technique to incorporate metadata available for the analyzed subjects. Data like genotype, or Aβ_{42} can give additional information beyond MR appearance and can be expected to better model the resulting low-dimensional representation. After finding such a low-dimensional representation it can be used to perform classification between clinical subject groups. The presented results show a classification accuracy that compares favorably to established neuroanatomical biomarkers and a significant improvement with the incorporation of non-imaging metadata.

Chapter 7 presents different ways to model longitudinal brain development in a low-dimensional manifold representation. Classification results improve significantly when using longitudinal information.

1.4.4 Comprehensive analysis of the developed biomarkers

Finally, Chapter 8 presents a comprehensive analysis on the ability of the proposed biomarkers in combination with other measures extracted from MRI to discriminate between clinical subject groups. A clear improvement in classification accuracy is observed for a combination of several biomarkers.

Chapter 2

Background

This chapter gives an overview on some of the most important developments in the two main fields, this thesis deals with. While individual chapters in this work give an introduction to the topic covered and place it within the context of existing methods, this chapter gives an introduction to the research area in a broader sense. The first part describes some of the most prominent methods for an automated segmentation of brain structures and atrophy measurement. In the second part, established methods for dimensionality reduction are presented.

2.1 Brain atlases

Brain atlases are defined by anatomical labels in a stereotaxic space, i.e., a standardized coordinate system that establishes a mapping from the voxel in one brain to the corresponding voxel in a second brain. Aligning an unseen image with the defined labels in the reference space allows to use this prior knowledge when processing the unlabeled image. A distinction can be made between probabilistic atlases that give at every voxel a probability of observing a particular structure, and atlases that give the manual labeling of an individual brain image.

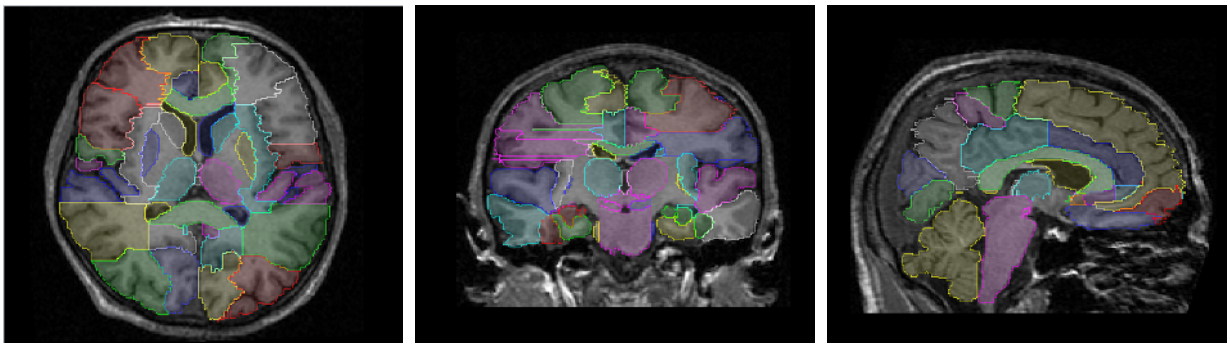
One of the first printed atlases that describe relations between different brain structures in a common space is the Talairach atlas [131] presented in 1967. First digital 3D atlases were designed in the 1980's, e.g., [11]. Most of the early brain atlases

were based on the manual labels on a small number of subjects. Probabilistic atlases developed later are, due to the time intesive labeling, mostly based on automated segmentation, e.g., [33]. A set of 30 manually labeled brain atlases that is used in this work is described in the next section.

2.1.1 Hammers brain atlases

In this thesis, a set of 30 brain atlases is used, each being manually delineated into 83 anatomical structures [67, 64][1]. The MR images used for atlas creation were acquired from young healthy subjects (age range 20-54, median age 30.5 years). Information on MR acquisition and a definition of the 83 delineated structures is given in Appendix B.

Figure 2.1.1 shows the manual segmentation overlaid on one of the 30 atlas MR images.



(a) Transverse (b) Coronal (c) Sagittal

Figure 2.1: Manually delineated structures on a brain atlas

2.2 Atlas based brain segmentation

A straight forward use of (manually) labeled atlas images is to transform them to the coordinate system of an unseen image and use the label maps to obtain the desired segmentation in target space. Early work in atlas based brain segmentation has been published by Collins et al. [32] and Christensen et al. [27] in the mid-1990's. In this

[1] www.brain-development.org

work, a single atlas image is nonlinearly aligned with a target image and the resulting transformation is used to propagate the structural label maps into target space. Figure 2.2 illustrates the concept of atlas-based segmentation.

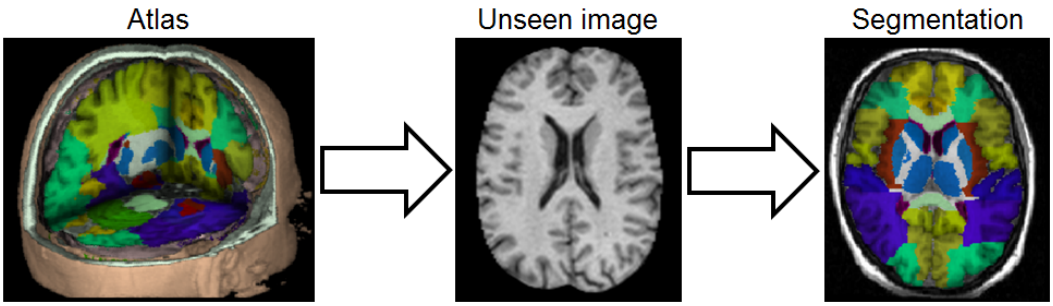


Figure 2.2: Atlas based brain segmentation: an atlas image is registered with the unseen image and atlas labels are propagated.

Such approaches crucially depend on the alignment of the atlas to the target image. The resulting segmentation fails in areas where the underlying registration fails. There are two general directions of research to overcome this limitation. In the first direction, a single individual atlas or a probabilistic atlas are used in combination with an intensity model to define the final segmentation. The second direction proposes to register multiple labeled atlases and to use techniques from machine learning to obtain a final segmentation from the individual atlas labels. In this thesis, methods are proposed that follow recent attempts to combine both conceptual directions.

In the reminder of this section, an overview on the development of both sketched paths is presented.

2.2.1 Atlas segmentation incorporating intensity modeling

Early work combining atlas-based brain segmentation with an intensity model is the tissue classification framework proposed by van Leemput et al. [142]. In this work, probabilistic brain atlases for the three tissue classes white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF) are used to initialize an expectation maximization (EM) framework to segment the image. Initialized from the probabilistic atlases, model parameters for Gaussian distributions describing the three tissue classes are

optimized with convergence of the algorithm. The method automatically corrects for MR intensity inhomogeneities and performs Markov random field based regularization of the segmentation to achieve a smooth labeling.

Building the basis for the widely used Freesurfer segmentation tool[2], Fischl et al. [53] presented a method that uses a probabilistic atlas in combination with an intensity model to segment 37 anatomical structures in addition to the three tissue classes. In this work, Gaussian intensity distributions for the structures of interest are learned from a set of labeled images. A probabilistic atlas build from all reference segmentations is then aligned with an unseen target image to give a spatial prior for the different structures. Following initialization, an MRF is defined on the target image with the data term being defined by the spatial prior together with the intensity model and smoothness constraints enforcing a consistent segmentation.

Ashburner and Friston [7] propose a *unified segmentation* framework that combines registration to a template space with tissue classification at the same time. In registration algorithms like the one used in the popular SPM software package[3] that are driven by a tissue classification, a combined approach is expected to improve results for both tasks. The objective function optimized in this work employs a mixture of Gaussians (MOG) model to represent the three tissue classes that accounts for smooth intensity variations caused by MR inhomogeneity. A deformable spatial prior incorporated into the objective function allows to include registration to a standard space into the model. This joined objective function is then optimized using an Iterated Conditional Modes (ICM) approach.

A hierarchical model for brain segmentation into tissue classes and anatomical structures has been proposed by Pohl et al. [113]. In a top to bottom approach, a subdivision of the brain according to prior information is performed along a tree structure representing the brain at different resolutions. In this tree, "brain", e.g., is a parent node of "WM", "GM", "CSF". The segmentation at every level is performed by an EM-based algorithm similar to the one for tissue class segmentation described above

[2]http://surfer.nmr.mgh.harvard.edu/
[3]http://www.fil.ion.ucl.ac.uk/spm/

[142]. While such a hierarchical approach allows to easily subdivide the segmentation problem, it implicitly assumes a perfect segmentation in the higher levels, making it impossible to recover from segmentation errors in later stages.

More recently, specialized methods to tackle more specific problems have been published. A method that delivers a state of the art automated segmentation of hippocampus and amygdala has been proposed by Chupin et al. [31, 30]. Here, an initial segmentation of both structures obtained from a registered probabilistic atlas and estimated intensity models is iteratively deformed in a topology preserving manner. Neuroanatomical landmarks not only derived from hippocampus and amygdala but also from neighboring structures are used to define a Markovian energy function following empirical descriptions of patterns in brain anatomy. Hippocampus and amygdala bordering regions are then deformed in an alternating fashion, optimizing the Markovian energy function using the ICM algorithm.

2.2.2 Multi-atlas segmentation

The idea behind multi-atlas segmentation is to make atlas-based segmentation more robust against errors in the registration of an atlas image by registering multiple atlases with the target image before obtaining a consensus segmentation from the individual labels. This concept is illustrated in Figure 2.3.

In [116], Rohlfing et al. show on images of bee brains how the segmentation accuracy can be improved by registering multiple atlases instead of a single atlas. Using an approach from pattern recognition, "Vote Rule" decision fusion is carried out, assigning to each voxel the label that receives the most "votes" from the individual atlases. This framework has been shown to significantly improving atlas segmentation and was successfully applied to human brain segmentation with the 30 atlas images described in Section 2.1.1 by Heckemann et al [73].

Different strategies to select suitable atlases in an atlas segmentation scheme, in particular multi-atlas segmentation, have been proposed. The STAPLE algorithm presented by Warfield et al. [148] describes a general framework to give a probabilis-

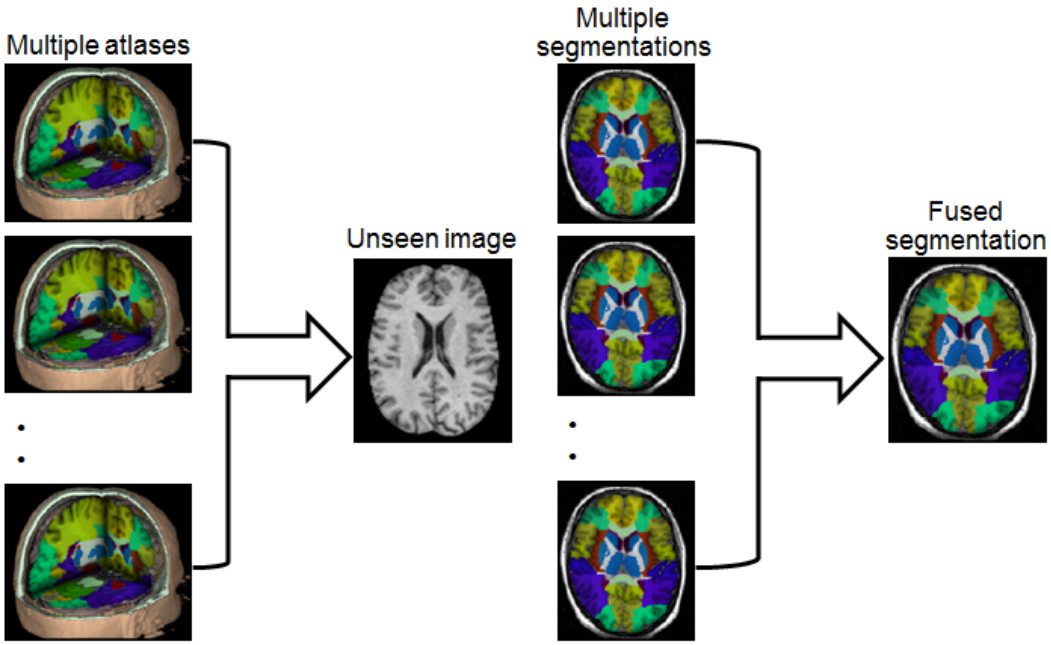


Figure 2.3: Multi-atlas brain segmentation. Multiple atlas images are registered with the unseen images and all atlas labels are propagated to the target. The final segmentation is obtained from fusing the individual segmentations.

tic estimate of a segmentation by weighting a number of individual segmentations while considering the performance of every individual segmentation. An EM framework is described that iteratively estimates the true segmentation by weighting all individual segmentations and then updating the final segmentation estimate based on these weightings. Aljabar et al. [1] propose a different strategy that a-priori selects a set of atlases from an atlas pool before performing multi-atlas segmentation with majority vote [116, 73]. Based on simple intensity-based metrics or subject-based meta-information, all available atlases are ranked according to their suitability for a given query image. By using the top-ranked images as atlases, registration error can be kept to a minimum resulting in an optimized segmentation.

More sophisticated atlas-selection techniques allowing a local assignment of suitable atlases have been proposed recently [4, 119].

2.2.3 Multi-atlas segmentation incorporating intensity modeling

Recent work proposes a combination of multi-atlas segmentation and intensity-based refinement. van der Lijn et al. [140] propose to generate a target-specific probabilistic hippocampus atlas by registering multiple atlas images. The obtained spatial prior is combined with a previously learned intensity model for the hippocampus to define an MRF-based energy function which is then optimized using graph cuts. Chapter 3 of this thesis presents a fully automated extension of this framework that takes advantage of multiple defined brain structures. Lötjönen et al. [102] use the popular EM algorithm described for tissue class segmentation above [142] to refine the segmentation estimate obtained from multi-atlas segmentation. A comparison of this algorithm with the graph-cut based approach presented in this thesis (Chapter 3) has been published in [104].

2.3 Atrophy measurement

Intra-subject brain changes over time have been shown to provide a more accurate biomarker for AD than cross-sectional differences.

Several methods to accurately measure structural volume changes in brain images have been developed. Freeborough and Fox [57] proposed the boundary shift integral (BSI) which quantifies structural volume change between rigidly registered repeat MR scans. Based on the segmentation in baseline and follow-up scan, the shift of an object boundary is measured. Structural volume loss is then estimated by integrating over the intensity differences in the shifted area. Differences are only evaluated over a defined intensity window to get more robust against segmentation errors. While initially based on whole-brain atrophy and manual segmentation of baseline and follow-up scan, a recent publication applies it successfully on hippocampal atrophy in the ADNI database using a fully automated multi-atlas segmentation approach [97].

Another technique based on the registration between follow-up and baseline image

is Structural Image Evaluation, using Normalization, of Atrophy (SIENA) [127]. This technique starts with extracting the brain at baseline and follow-up using a tessellated surface mesh. After co-registering both images, a combined brain mask is produced. Using a gradient-based edge detector, the method then finds all brain surface points in both images to estimate the motion of each point over time. Matching the gradient points in both images, finally allows to measure atrophy on a voxel basis.

Deformation-based morphometry (DBM) [8] was originally proposed as a method to measure inter-subject differences from the deformation fields obtained from non-rigidly aligning a set of subjects to a template space. In the original publication, non-rigid deformations are parametrized by a linear combination of discrete cosinus transform (DCT) basis functions [5]. Analyzing the coefficients of individual deformation fields allows to identify anatomical group differences resulting in systematically different deformation fields. Freeborough and Fox [58] propose to model intra-subject brain deformations by inspecting the deformation fields obtained from registering a follow-up scan to its baseline using a fluid registration algorithm [32]. Determining the Jacobian matrix of the deformation field at voxel level allows to measure whether there is expansion (Jacobian determinant > 1) or contraction (Jacobian determinant < 1). Integrating the Jacobian determinant over a region of interest gives an estimate of atrophy in this region. With nonrigid registration using free-form deformation based on B-splines [118], this technique was successfully applied to measuring cerebral atrophy in MR brain images [17]. A cross-sectional analysis of the ADNI database based on DBM was recently published by Hua et al. [77].

More recent approaches to measure atrophy include a method developed by Thompson et al. [134] that uses 3D surface meshes based on manual segmentations at baseline and at follow-up to extract 3D maps of structural development. This method was applied to measure hippocampal atrophy in ADNI with an automated hippocampal segmentation method [109]. Xue et al. [156] present a framework to measure atrophy from the segmentation of baseline and follow-up scan. This work shows how a more accurate measurement of atrophy can be achieved by incorporating spatial constraints

into a 3D segmentation method for the simultaneous analysis of longitudinal images.

2.4 Manifold learning

The second part of this thesis presents different techniques to apply dimensionality reduction techniques for the extraction of biomarkers. In this section, different widely used techniques for dimensionality reduction are presented. The overview given follows the detailed description of manifold learning techniques given by van der Maaten et al. [141].

A set of images $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\} \in \mathbb{R}^D$ is described by N images $\mathbf{x}_i$, each being defined as a vector of intensities, where D is the number of voxels per image or region of interest (typically $D >$1,000,000 for brain MR images). Assuming $\mathbf{x}_1, ..., \mathbf{x}_N$ lie on or near an d-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^D$, it is possible to learn a low dimensional representation $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$ with $\mathbf{y}_i \in \mathbb{R}^d$ of the input images in $\mathcal{M}$ of the input images.

In many of the techniques described, a matrix is typically used to represent the relations between pairs of data items, which, for the purpose of this thesis, can be assumed to be images. The matrix in turn, may be viewed as representing a graph to model the data in which each node is an image and the weight of each edge denotes the similarity or dissimilarity between the image pair it joins. A broad distinction can be made between methods that use a complete graph to model the relations among the data and those methods that use a sparser representation with a smaller number of edges, restricted to local neighborhoods. All the methods below seek to optimize some form of objective function via the matrix representation. The techniques are described as spectral as the optimization is often carried out using the eigenvalue-eigenvector structure of the associated matrix.

A schematic overview of manifold learning techniques is given in Figure 2.4.

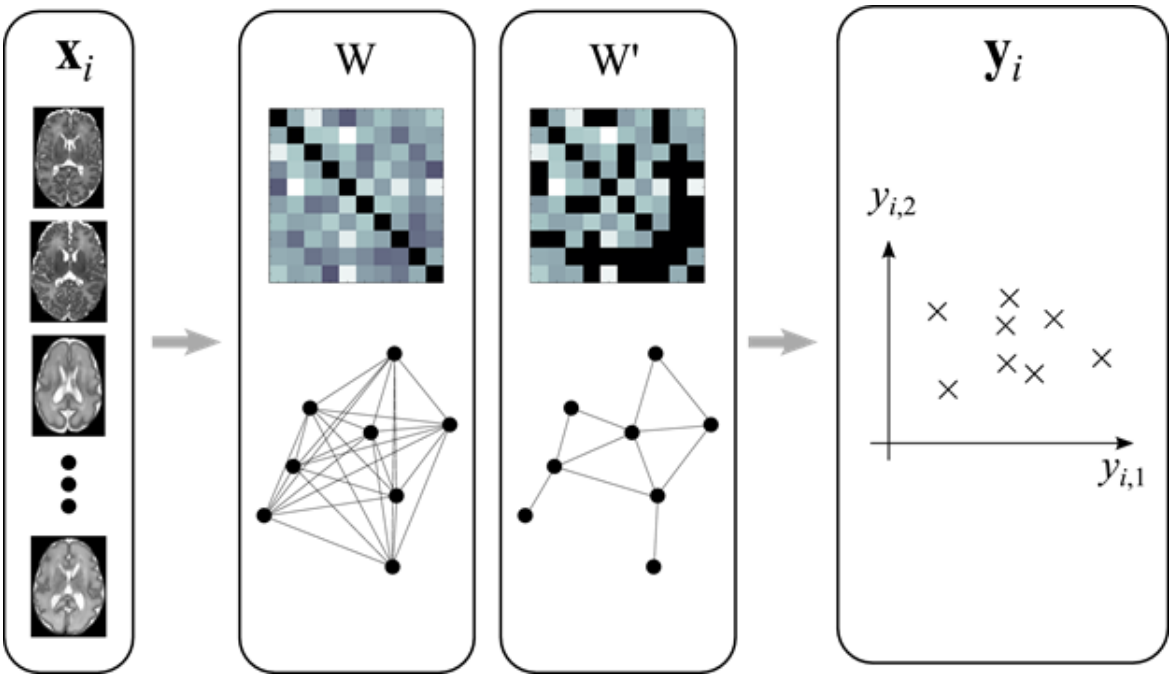


Figure 2.4: In the schematic illustration above, the images $\mathbf{x}_i$, $1 \leq i \leq N$ (left) are compared in pairs and measures of similarity or distance between them are obtained. The measures define a $N \times N$ matrix representing the edge weights in a graph representation of the data. The graph/matrix representation may be either full (dense, $\mathbf{W}$ above) or sparse ($\mathbf{W}'$), illustrations of both cases are shown above. Typically, the eigenvalue-eigenvector structure of the matrix (or of a matrix derived from it) is used to derive a coordinate representation for an embedded manifold representation $\mathbf{y}_i$ of the original data. The first two dimensions of $\mathbf{y}_i$ are schematically shown above.

2.4.1 Dense spectral techniques

This section describes dense techniques for manifold learning, using a full matrix of pairwise relations to learn the low-dimensional representation. The full data matrix $\mathbf{X}$ is constructed so that its i-th row is the data item $\mathbf{x}_i$ and the low dimensional representation $\mathbf{Y}$ similarly contains $\mathbf{y}_i$ as its rows.

PCA

Principal Component Analysis (PCA) [84] is a popular and widely used linear dimensionality reduction technique. PCA aims to describe as much of the variance in the data using only a few principal components. The problem is described as finding the

linear mapping function $\mathbf{M}$ that optimizes the objective function

$$\max_{\mathbf{M}} \operatorname{trace}\left(\mathbf{M}^T \operatorname{cov}(\mathbf{X})\,\mathbf{M}\right) \tag{2.1}$$

where $\operatorname{cov}(\mathbf{x})$ is the sample covariance matrix of $\mathbf{X}$. The linear mapping is defined by the first d eigenvectors of the eigenproblem

$$\operatorname{cov}(\mathbf{X})\,\mathbf{M} = \lambda \mathbf{M}. \tag{2.2}$$

From this, the mapping into low-dimensional space is defined as $\mathbf{Y} = \mathbf{XM}$

Kernel PCA

Kernel PCA [123] is a nonlinear extension of classic PCA. A kernel matrix $\mathbf{K}$ is defined from the data points in D-dimensional space with

$$\mathbf{k}_{ij} = \kappa(\mathbf{x}_i, \mathbf{x}_j), \tag{2.3}$$

where κ can be any function that results in a positive-semidefinite $\mathbf{K}$. A centering operation is performed subsequently to make the defined features zero-mean and computing the d principal eigenvectors $\mathbf{v}_i$ and eigenvalues λ_i of $\mathbf{K}$, leads to the eigenvectors $\mathbf{a}_i$ of the associated covariance matrix:

$$\mathbf{a}_i = \frac{1}{\sqrt{\lambda_i}}\mathbf{v}_i. \tag{2.4}$$

The low-dimensional embedding of image $\mathbf{x}_i$ is then defined as

$$\mathbf{y}_i = \left\{ \sum_{j=1}^{N} \mathbf{a}_1^{(j)} \kappa(\mathbf{x}_j, \mathbf{x}_i), ..., \sum_{j=1}^{N} \mathbf{a}_d^{(j)} \kappa(\mathbf{x}_j, \mathbf{x}_i) \right\} \tag{2.5}$$

where $\mathbf{a}_i^{(j)}$ is the j-th entry of vector $\mathbf{a}_i$.

MDS

Multidimensional scaling (MDS) [36] is a linear technique closely related to PCA. It is based on a distance matrix $\mathbf{D}$ with d_{ij} representing the distance between two high-dimensional data items $\mathbf{x}_i$ and $\mathbf{x}_j$. MDS seeks to find the low-dimensional representation that best preserves the pairwise distances in the high-dimensional space. This is carried out by minimising the objective function

$$\phi(\mathbf{Y}) = \sum_{ij} \left(d_{ij}^2 - \|\mathbf{y}_i - \mathbf{y}_j\|^2 \right) \tag{2.6}$$

with $\|\mathbf{y}_i - \mathbf{y}_j\|$ being the distance between two datapoints in d-dimensional space, d$\ll$D. The optimal embedding for this objective function can be obtained through a singular value decomposition of the Gram matrix $\mathbf{K} = \mathbf{X}\mathbf{X}^T$ which may be derived from the distance matrix $\mathbf{D}$.

Isomap

Isomap [133] is a nonlinear embedding technique that builds upon the MDS approach. In Isomap pairwise distances d_{ij} are not measured directly between data items $\mathbf{x}_i$ and $\mathbf{x}_j$ but on a neighborhood graph G connecting all N data items. This graph is defined by either connecting every data item $\mathbf{x}_i$ to its k closest neighbors or to all subjects within some fixed radius ϵ. After constructing G, the distances d_{ij} are estimated as the shortest path distances d_{ij}^G within the graph. The final embedding coordinates $\mathbf{y}_i$ are obtained by applying classical MDS to the distance matrix $\mathbf{D}^G = \left\{ d_{ij}^G \right\}$.

2.4.2 Sparse spectral techniques

In this section, some of the available sparse techniques for manifold learning are described that focus on retaining the local similarities measured in the input space.

LLE

A low-dimensional manifold constructed with Locally Linear Embedding (LLE) [117] aims to preserve the local neighborhoods of the high-dimensional data in the learned low-dimensional space. The method assumes a locally linear relationship between neighboring data points. The idea is to represent every data item $\mathbf{x}_i$ as a weighted combination of its k closest neighbors in the high-dimensional space. This defines a set of weights w_{ij} for the k neighbors of $\mathbf{x}_i$ and the aim is to find a low-dimensional representation $\mathbf{y}_i$ that respects this weighting. The LLE objective function is defined as

$$\phi(\mathbf{Y}) = \sum_i \left\| \mathbf{y}_i - \sum_{j=1}^{k} \mathbf{w}_{ij}\mathbf{y}_{ij} \right\|^2 \text{ subject to } \left\|\mathbf{y}^{(k)}\right\|^2 = 1. \tag{2.7}$$

With the sparse weight matrix $\mathbf{W}$, the embedding is obtained from the d eigenvectors corresponding to the smallest nonzero eigenvalues of $(\mathbf{I}-\mathbf{W})^T(\mathbf{I}-\mathbf{W})$.

Hessian LLE

Using the same concept of local linearity as LLE, Hessian LLE [43] minimizes the curvature of the high-dimensional manifold when learning the low-dimensional representation. The method enforces local isometry between the distances in both spaces. Applying PCA to every datapoint $\mathbf{x}_i$ and its k nearest neighbors gives an approximation of the local tangent space at every data point. The mapping function $\mathbf{M}$ obtained from the d principal components at every point $\mathbf{x}_i$ is then used to give an estimator for the Hessian $\mathbf{H}_i$ of the manifold at that data point [43]. From the Hessian estimators in tangent space, a matrix $\mathcal{H}$ is constructed with entries

$$\mathcal{H}_{lm} = \sum_i \sum_j \left((\mathbf{H}_i)_{jl} \times (\mathbf{H}_i)_{jm} \right) \tag{2.8}$$

The eigenvectors that correspond to the d smallest eigenvectors of $\mathcal{H}$ are used to define the low-dimensional embedding $\mathbf{Y}$ that minimizes the curvature of the manifold.

Laplacian eigenmaps

Laplacian eigenmaps [14] aims to learn a manifold representation that preserves a set of similarities in a local neighborhood for the high-dimensional data. Weights w_{ij} are defined as the similarities between subjects within a local neighborhood and set to zero for all other pairings. Similarities can be derived from distances d_{ij} using a heat kernel such as

$$w_{ij} = e^{-\frac{d_{ij}^2}{t}}. \tag{2.9}$$

The LE embedding is obtained by minimizing the objective function

$$\phi(\mathbf{Y}) = \sum_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|^2 w_{ij} = 2\mathbf{Y}^T\mathbf{L}\mathbf{Y} \tag{2.10}$$

where $\mathbf{L} = \mathbf{D} - \mathbf{W}$ is the graph Laplacianmatrix which is derived from the weight matrix $\mathbf{W}$ and the diagonal degree matrix $\mathbf{D} = \sum_j w_{ij}$. The LE objective function is optimized under the constraint that $\mathbf{y}^T\mathbf{D}\mathbf{y} = 1$ which removes an arbitrary scaling factor in the embedding and prevents the trivial solution where all $\mathbf{y}_i$ are zero. The $\mathbf{y}_i$ that optimize the objective function are defined by the eigenvectors corresponding to the smallest nonzero eigenvalues of the generalized eigenvalue problem $\mathbf{L}\mathbf{d} = \lambda\mathbf{D}\mathbf{d}$.

2.4.3 Application of manifold learning

With different manifold learning techniques being tailored to address different dimensionality reduction problems, each involving several parameters to set, an application is not always straight forward. Depending on the expected underlying space, a choice has to be made for a linear or nonlinear technique. The nonlinear techniques described above minimize an objective function based on a local neighborhood in the input space. A crucial parameter with these methods is k, the number of neighbors considered for every subject and therefore defining the expected degree of nonlinearity. Another important choice has to be made regarding the input measure, whether it is based on a distance metric or a pure similarity measure. While a distance measure can

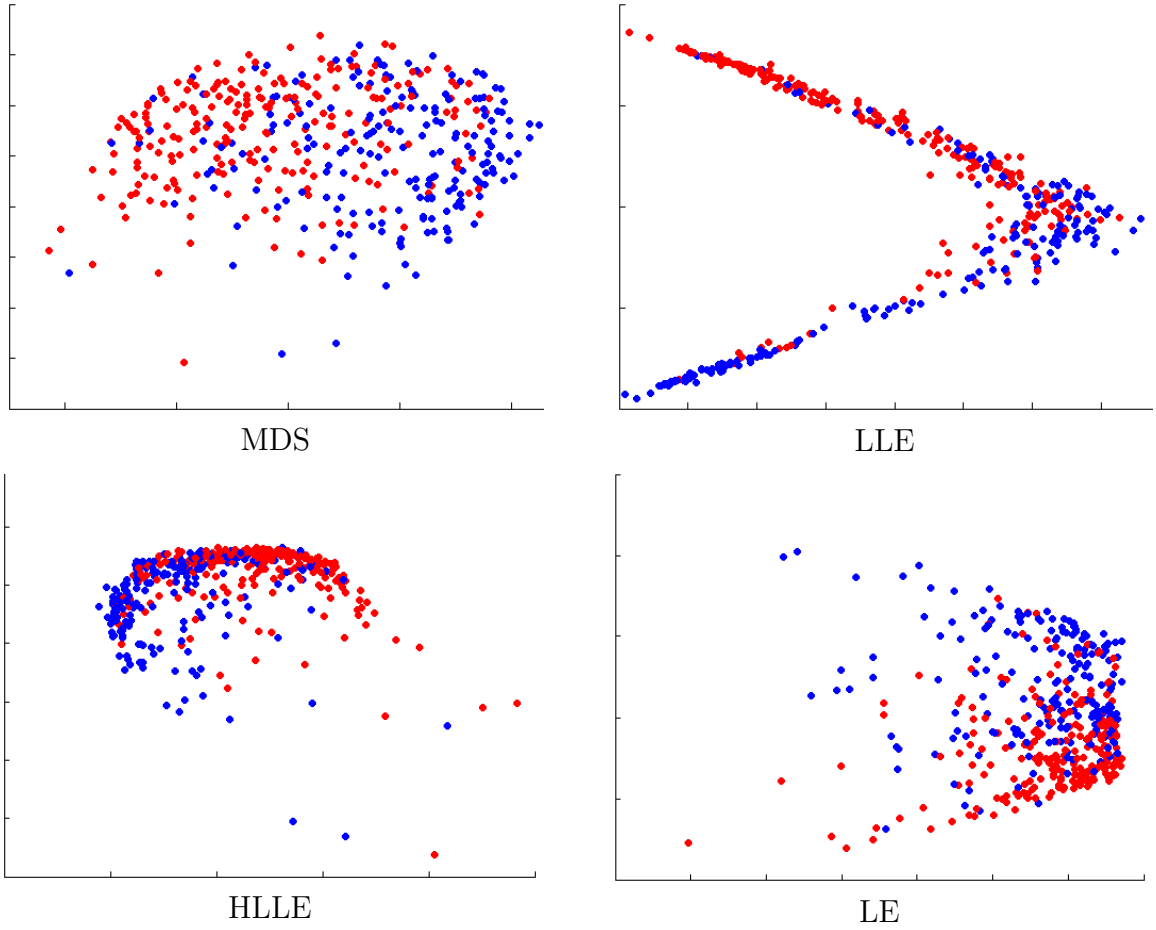


Figure 2.5: 2D embeddings obtained with different dimensionality reduction techniques to 167 images acquired from AD patients (blue) and 231 images from healthy controls (red).

be converted into a similarity (e.g. by using the heat kernel given in Equation 2.9), this conversion brings an additional parameter, here t, with it. Equally, a similarity-based measure can be converted into a distance measure only under certain assumptions. The application of an embedding technique that readily deals with the available input measure is therefore recommended. The input measure used with manifold-learning in this thesis is derived from intensity-similarities. With Laplacian eigenmaps being able to readily deal with similarities, it is used for all applications described in the following.

As an illustration, results obtained from different embedding techniques are displayed in Figure 2.5. The four plots show manifold embedding coordinates obtained using MDS, LLE, HLLE and Laplacian eigenmaps (LE). For 167 images acquired from

subjects with Alzheimers disease and 231 images from healthy controls, the pairwise similarity measure s_{ij} is defined as the cross correlation between each pair of images $\mathbf{x}_i$ and $\mathbf{x}_j$. For the distance-based learning methods, the similarity s_{ij} is transformed into a distance d_{ij} with $d_{ij} = 1 - s_{ij}$. A neighborhood size of $k = 15$ is used for all sparse methods. The first two dimensions of the resulting embedding coordinates are plotted for each of the different methods (AD subjects are plotted in blue and healthy controls in red).

Chapter 3

Automated intensity-refinement with multi-atlas label propagation

This chapter is based on:

Robin Wolz, Paul Aljabar, Rolf A. Heckemann, Alexander Hammers, Daniel Rueckert. "Segmentation of Subcortical Structures and the Hippocampus in Brain MRI using Graph-Cuts and Subject-Specific A-Priori Information". *ISBI 2009*, Boston, USA, Juli 2009.

Abstract

This chapter describes a general framework for the segmentation of subcortical structures and the hippocampus in magnetic resonance brain images based on multi-atlas label propagation and graph cuts. The label maps obtained from multi-atlas segmentation are used to build a subject-specific probabilistic atlas of a structure of interest. From this atlas and an intensity model estimated from the unseen image, a Markov random field-based energy function is defined and optimized via graph cuts. Compared to a previously proposed approach, this method does not rely on manual training of the intensity model. It is applied to five subcortical structures and the hippocampus. The presented method is used to segment the hippocampus on 60 ADNI images and an average overlap (Dice coefficient) of 0.86 was obtained with reference segmentations.

3.1 Introduction

The accurate and robust segmentation of subcortical brain structures and the hippocampus in magnetic resonance images is an increasingly important step in the diagnosis of Alzheimer's disease. Although much research has been published in this area [53, 113, 73, 122, 10], no method has established itself in routine clinical use. One well-validated approach relies on combining the segmentations obtained from non-rigidly aligning multiple manually labeled atlases with the target image [73]. The final label at each voxel is determined by applying vote-rule decision fusion. This method makes no use of the target intensity information. Considering such information, however, potentially results in further improvements to the quality of multi-atlas segmentation.

Combining prior knowledge of the intensity and spatial distribution of an object of interest in the contextual framework of a Markov random field (MRF) is an established technique for brain segmentation (e.g. [53, 122, 140]). In these approaches spatial information in the form of a probabilistic atlas and an estimation of the probability distribution of the target structure's intensities are used to formulate an energy function. Introduced by Greig et al. [65] and proposed as a generic method for finding the global optimum for labeling tasks in computer vision by Boykov et al. [18], graph cuts have been widely used for optimization in this area.

Recently, two brain segmentation methods based on MRFs and graph cuts have been introduced: Song et al. [128] proposed a method for tissue class segmentation of 2D MR images. Their spatial prior is defined as a probabilistic atlas that is affinely registered to the target image. The intensity distributions of white matter (WM), gray matter (GM), and cerebrospinal fluid (CSF) are modeled using Gaussian distributions. Another promising approach, proposed by van der Lijn et al. [140] for segmenting the hippocampus, can be considered an extension of the multi-atlas segmentation approach of [73] and tackles the previously described problem as follows: instead of directly fusing the individual segmentations obtained from registering multiple atlases to the target image, they are used to build a probabilistic atlas which is combined with statistical intensity models for foreground and background to formulate an energy

function to be minimized. A limitation of this method is the reliance on a strictly controlled training of its statistical intensity model where a Gaussian distribution for the hippocampus and a Parzen estimate of the background distribution are defined on the manually labeled atlas images. This approach requires the use of identical MR sequences for the atlas (training images) and target (subject images).

This chapter describes a generalized framework for the segmentation of subcortical brain structures and the hippocampus in MR images which overcomes these problems by directly estimating the Gaussian distribution for the foreground from the target image. Furthermore, a spatially varying mixture of Gaussians (MOG) model for the background is used in order to better model the different background parts surrounding a structure of interest. The method is extended to five subcortical structures and the hippocampus and evaluated on 60 ADNI images.

3.2 Method

The task of segmenting an image I into structures of interest can be described as assigning a label $f_p \in Ł$ to each voxel $p \in I$. A MRF-based energy function can be formulated as

$$E(f) = \sum_{p \in I} D_p(f_p) + \lambda \sum_{\{p,q\} \in N} V_{p,q}(f_p, f_q), \tag{3.1}$$

where N is a neighborhood of voxels and f is the labeling of I [18]. The data term $D_p(f_p)$ measures the disagreement between a probabilistic model and the observed data. $V_{p,q}(f_p, f_q)$ is a smoothness term penalizing discontinuities in N. The parameter λ weights the influence of the data term and the smoothness term. For the evaluation described in Section 3.3 it was set to an empirically determined value of $\lambda = 0.5$.

To optimize Equation (3.1) with graph cuts, a graph $G =< V, E >$ with a node $v \in V$ for each voxel p is defined on image I. Its edges $e \in E$ consist of connections between each node v and two terminal nodes s, t as well as connections between neighboring voxels. The terminals s and t represent the two labels describing foreground and

background. By determining an s-t cut on G, the desired segmentation can be obtained [18]. The data term in the MRF model defines the weights of the edges connecting each node with both terminals and the smoothness term encodes the edge weights of neighboring nodes. The segmentation with graph cuts is illustrated in Figure 3.1.

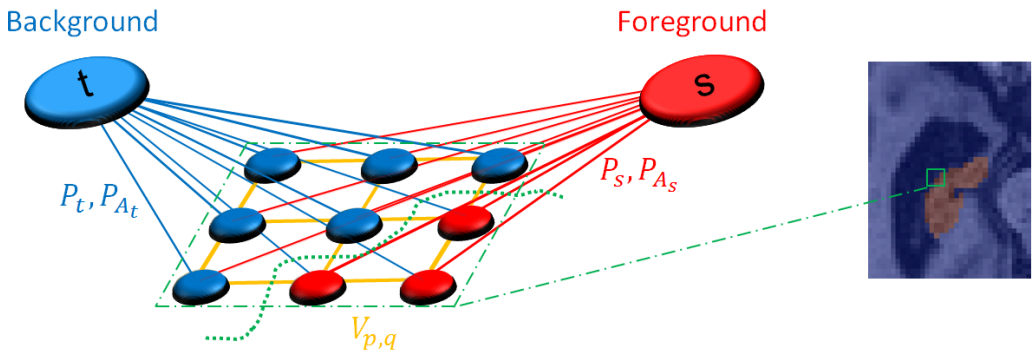


Figure 3.1: Segmentation with graph cuts. A graph is defined on the target image in which every voxel is represented by a node. Source and sink nodes represent foreground and background and weights from every node to source and sink are defined according to the energy model. Edges connecting neighboring nodes enforce smoothness. The final segmentation is obtained by finding the minimum cut of the defined graph.

To guarantee a global optimum, this segmentation based on graph cuts can only be applied to a binary segmentation problem where an image is segmented into foreground and background. To segment multiple structures, the algorithm can be applied for each structure S_i independently before consolidating the individual segmentations in a final step. (Equivocal voxels are labeled according to the spatial prior introduced in Section 3.2.1).

3.2.1 Estimation of a subject-specific data term

The weights of the edges connecting each node with the terminals are determined from a spatial prior and a model of the intensity distribution of the structure of interest. To estimate the corresponding parameters as accurately as possible, both models are derived from the unseen target image.

Spatial prior

Various authors have used prior spatial knowledge in the shape of a probabilistic atlas for MRF-based brain segmentation (e.g. [122, 142, 128, 140]). While most of these approaches rely on affinely aligning a fixed probabilistic atlas for tissue classes or individual structures, van der Lijn et al. [140] use the propagated labels from multi-atlas segmentation [73] to build a subject-specific probabilistic atlas directly in the coordinate system of the unseen image. Building an atlas from multiple registrations compensates for errors in the constituent atlases and registrations. Here, a similar approach is proposed using a non-rigid registration method [118] to align all N atlases with the target image. The parameter settings for image alignment are based on the well-evaluated procedure described in [73]. By applying the resulting transformations T^j to each label set f^j, each atlas is warped to the target image's coordinate frame.

For each voxel p, the prior probability of its label being f_i is therefore

$$P_A(f_i) = \frac{1}{N} \sum_{j=1,\ldots,N} \begin{cases} 1, & f_i = f_i^j \\ 0, & \text{else} \end{cases} \tag{3.2}$$

P_A defines the spatial prior contribution to the data term in the graph cuts model.

Intensity model

The intensity prior for tissue classes or specific structures is usually modeled by a Gaussian probability distribution. The main challenge is the accurate and robust estimation of its parameters. In [142], van Leemput et al. describe an expectation-maximization based method to successively improve an initial estimate of the parameters of tissue class distributions. For the hippocampal segmentation proposed in [140], the parameters of the Gaussian distribution are estimated *a priori* from manually labeled training images, which restricts the method to test and training images with identical MR sequences. To arrive at a more generally applicable method, in this work the parameters of the Gaussian distribution of the structures of interest are directly estimated from the unseen target image. It is estimated from all those voxels which at least 95% of

the atlases assign to this particular structure. The intensity component of the source link weight for a given voxel p with intensity y_p and structure f_i is denoted by P_s and is estimated from the intensity distribution model, i.e. $P_s(p, f_i) = P(y_p|f_i)$.

For many subcortical structures, the background is not typically homogeneous. Therefore, it is meaningful to describe its intensity distribution by a multivariate model instead of a single Gaussian distribution. Van der Lijn et al. [140] proposed a Parzen window estimated from a manually outlined area around the hippocampus on training images. To enable a more robust approach that does not rely on manual training and to allow for a more detailed description by using different models for different parts of the background, a spatially varying mixture of Gaussians (MOG) model is used in this work. The MOG model is defined by the general Gaussian distributions of the three tissue classes based on the method described in [142] and the more precise distributions of the defined regions of interest (subcortical structures and hippocampus) based on the target specific atlas described above. When segmenting a particular structure i with label f_i, the Gaussian intensity distributions of all other structures with labels f_j, $j \neq i$ and of the tissue classes T_k, $k = 1, ..., 3$ are combined to estimate the probability of the voxel belonging to the background. This is carried out using spatial priors for the structures (obtained as described above) and for the tissue classes (obtained from previously generated and non-rigidly aligned probabilistic atlases). The probability of a voxel being in the background with respect to structure i is estimated by:

$$\begin{aligned} P(y_p|f_{i,\text{back}}) &= (1 - \gamma_{\text{struct}}) \sum_{k=1,...,3} \gamma_k P(y_p|T_k) \\ &+ \gamma_{\text{struct}} \sum_{j=1,...,N, j\neq i} \gamma_j P(y_p|f_j), \end{aligned} \tag{3.3}$$

where γ_k is the tissue spatial prior, $\gamma_j = P_A(f_j)$ is the structure spatial prior and $\gamma_{\text{struct}} = \sum_{j=1,...,N, j\neq i} \gamma_j$. Equation 3.3 provides the intensity component of the edge weight from voxel p to the sink node t for the current structure, denoted by $P_t(p, f_i)$, i.e. $P_t(p, f_i) = P(y_p|f_{i,\text{back}})$

The intensity and spatial contributions, P_x, $x \in s, t$ and P_A, are combined to give

the data term that defines the edge weights connecting each node to the source s and sink t. It is defined as the log-likelihood:

$$D_p(f_i) = -\alpha \ln P_x(p, f_i) - (1-\alpha) \ln P_{A_x}(f_i) \tag{3.4}$$

With $P_{A_s}(f_i) = P_A(f_i)$ and $P_{A_t}(f_i) = 1 - P_A(f_i)$. The parameter α governs the influence of P_A and P_x on the final segmentation result.

3.2.2 Smoothness term

Following [128], a smoothness term based on intensity y as well as the intervening contour probabilistic map B (derived from the gradient image) are used to define the weights of edges connecting two neighboring voxels p and q:

$$\begin{aligned} V_{p,q}(f_p, f_q) &= c\left(1 + \ln\left(1 + \frac{1}{2}\left(\frac{|y_p - y_q|}{\sigma}\right)^2\right)\right)^{-1} \\ &+ (1-c)\left(1 - \max_{x \in M_{p,q}}(B_x)\right) \end{aligned} \tag{3.5}$$

where $M_{p,q}$ is a line joining p and q, and σ is the robust scale of image I [128]. The parameter c controls the influence of the boundary- and intensity based part and is empirically set to 0.5.

3.3 Data and Results

The method was evaluated on 60 T1-weighted 1.5T MR images from different subjects in the ADNI database described in Section 1.3 . The subjects in this study are classified into three groups: Alzheimer's patients (AD), patients showing mild cognitive impairment (MCI) and control subjects (controls). From each group 20 subjects were selected randomly. For each image a reference hippocampal segmentation was provided by ADNI (see Appendix A.1.1).

Two different sets of atlases were used for the segmentation. The first set consisted of 30 ADNI images with corresponding hippocampus labels as described in Section A.1.1. The subjects were different from those used for evaluation, and had been classified as AD, MCI, and controls (10 each). The first set of atlases was applied to compare the proposed method with the reference delineation for the hippocampus and multi-atlas segmentation. The second set of atlases consisted of the 30 Hammers atlases that are manually delineated into 83 structures and described in Section 2.1.1. This atlas set was used to segment the following structures for visual inspection: hippocampus, amygdala, putamen, thalamus, nucleus accumbens and caudate nucleus.

3.3.1 Comparison with manually labeled data

Table 3.1 shows the average overlap (similarity index, SI, or Dice coefficient) for the segmentation of the hippocampus for standard multi-atlas segmentation and the proposed method.

multi-atlas	0.842 ± 0.030 [0.739-0.894]
proposed method	0.860 ± 0.024 [0.787-0.897]

Table 3.1: Average SI overlap for hippocampus segmentation.

Figure 3.2 shows the difference between both methods for the 60 test images. This difference is statistically significant with $p < 0.001$ on Student's two-tailed paired t-test.

The improvements with the presented method are similar to those reported in [140], but are obtained without manually training the intensity models. To show the importance of such a sequence independent model, the proposed method was adapted to use a previously trained intensity model. The intensity distribution of the manually delineated hippocampi and the three tissue classes (WM, GM, CSF) was estimated on 10 MR ADNI-images which were acquired on the same scanner. Testing this model on 30 ADNI-images from *different* scanners at different sites, the average hippopcampal overlap was 0.851 compared to 0.848 for standard multi-atlas segmentation and 0.867 for the proposed method.

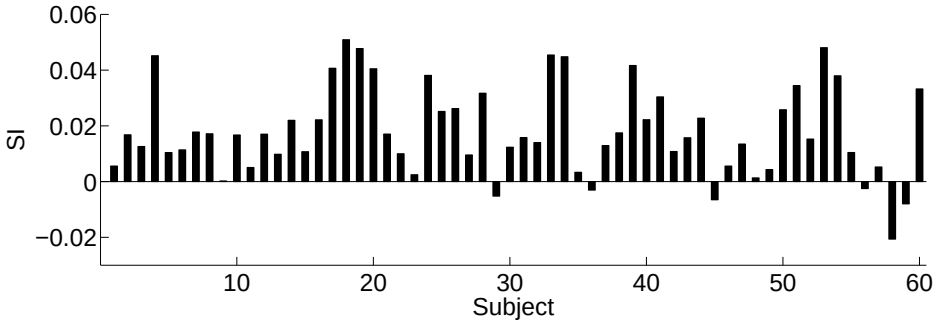
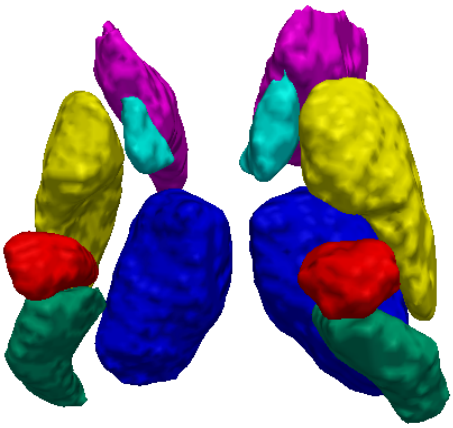


Figure 3.2: Difference between multi-atlas segmentation and the proposed method for the hippocampus segmentation in 60 test cases.

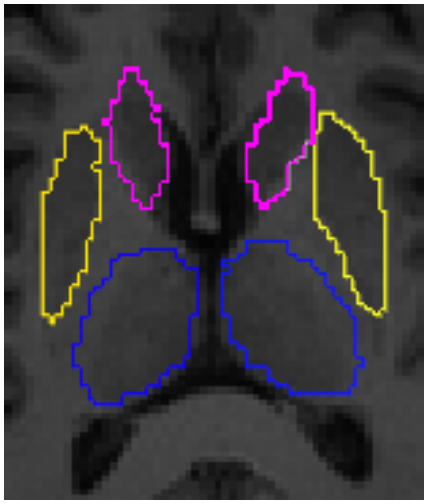
3.3.2 Visual inspection

Visual inspection of the segmentation results obtained from the second atlas set confirm the results described above and show improved segmentation results compared with standard multi-atlas segmentation.

Figure 3.3 shows the 3D-rendering for the 6 segmented structures and in a transverse slice the results for the thalamus, putamen and caudate.



(a) 3D rendering

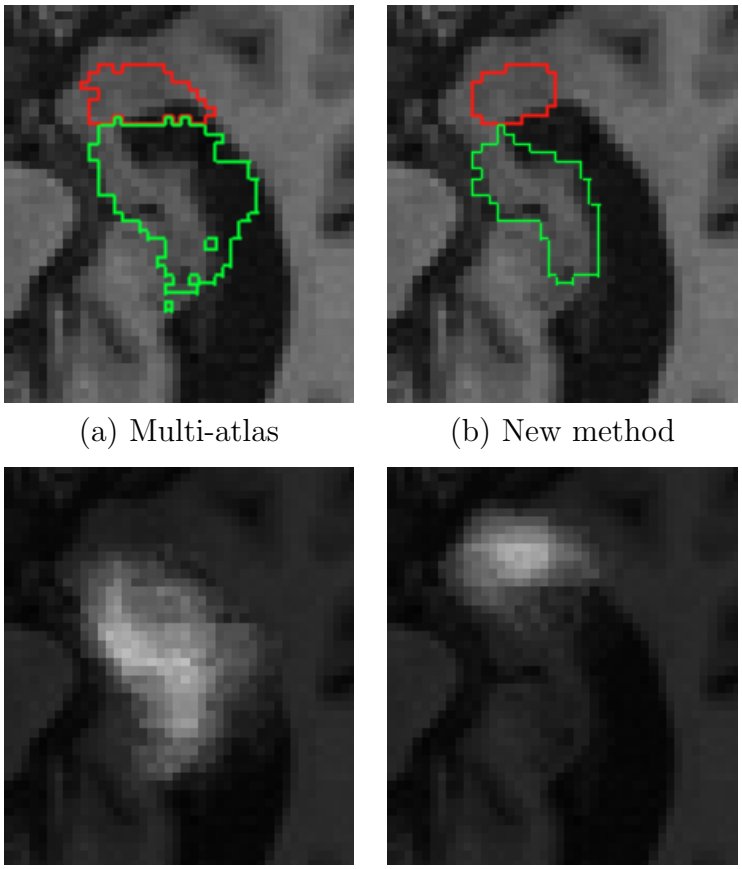


(b) Overview

Figure 3.3: (a) shows a 3D-rendering of the segmentation result for the proposed method for all structures: thalamus (blue), putamen (yellow), caudate (pink), hippocampus (green), amygdala (red) and nucleus accumbens (turquoise). (b): Transverse section showing segmentation outlines superimposed on an MR image.

In Figure 3.4 the results of multi-atlas segmentation, and the improved segmentation based on the proposed method are shown for the left hippocampus and amygdala for the MR image of an AD-patient. Furthermore examples of the subject specific atlas

which the proposed method builds on, are displayed. In this example, incorporating the automatically trained intensity model avoids substantial false-positive labeling.



(a) Multi-atlas (b) New method

(c) Atlases for hippocampus (d) Atlas for amygdala

Figure 3.4: (a) shows the segmentation results for multi-atlas segmentation (b): results for the proposed method. (c-d): Subject specific probabilistic atlases for hippocampus and amygdala (a higher intensity encodes a higher probability).

3.4 Conclusion

In this chapter, a method for subcortical brain segmentation in MR images based on subject-specific *a priori* information of spatial extent and intensity distribution of structures of interest was described. Label maps obtained from multi-atlas segmentation are used to generate a subject-specific probabilistic atlas. This atlas is paired with intensity models for both the foreground and the background to formulate an MRF-based energy function. In contrast to a previously proposed method, this algorithm does not rely on manual training of the intensity models. Therefore, this method is

more generally applicable as it is not tied to a specific MR sequence or contrast quality. A Gaussian distribution for the foreground model is directly estimated from the target image, while the background model is described by a mixture of Gaussians estimated from a tissue class segmentation, a subject-specific atlas and non-rigidly aligned atlases for tissue probabilities. The proposed method was evaluated on pathological image data from the ADNI study, increasing the SI overlap for the segmentation of the hippocampus significantly from 0.842 with standard multi-atlas segmentation to 0.860.

The following chapter describes a framework that uses the algorithm proposed here to propagate a set of atlases in a stepwise fashion to a diverse set of images, thereby reducing registration errors and increasing segmentation accuracy.

Chapter 4

LEAP: Learning Embeddings for Atlas Propagation

This chapter is based on:

Robin Wolz, Paul Aljabar, Joseph V. Hajnal, Alexander Hammers, Daniel Rueckert. "LEAP: Learning Embeddings for Atlas Propagation". *NeuroImage*, 49(2):1316-1325, 2010

Abstract

A framework for the automatic propagation of a set of manually labeled brain atlases to a diverse set of images is described. A manifold is learned that allows the identification of neighborhoods which contain images that are similar based on a chosen criterion. Within the new coordinate system, the initial set of atlases is propagated to all images through a succession of multi-atlas segmentation steps. This breaks the problem of registering images which are very "dissimilar" down into a problem of registering a series of images which are "similar". A set of 30 atlas images from young and healthy subjects is propagated to 796 images from elderly dementia patients and healthy controls from the ADNI study. The overlap of the automated hippocampus segmentation with reference labels is used for evaluation. An increasing gain in accuracy of the new method, compared to standard multi-atlas segmentation, is demonstrated

with a greater difference between atlas and image. The classification performance between clinical groups based on 83 structures, shows a significant improvement when using the described method compared to standard multi-atlas segmentation.

4.1 Introduction

Since brain anatomy varies significantly across subjects and can undergo significant change, either during aging or through disease progression, finding an appropriate way of dealing with anatomical differences during feature extraction has gained increasing attention in recent years. Amongst the most popular methods for dealing with this variability are atlas-based approaches: These approaches assume that the atlases can encode the anatomical variability either in a probabilistic or statistical fashion. When building representative atlases, it is important to register all images to a template that is unbiased towards any particular subgroup of the population [135]. Two approaches using the large deformation diffeomorphic setting for shape averaging and atlas construction have been proposed by Avants et al. [9] and Joshi et al. [85], respectively. Template-free methods for co-registering images form an established framework for spatial image normalization [129, 9, 158, 101, 15]. In a departure from approaches that seek a single representative average atlas, two more recent methods describe ways of identifying the modes of different populations in an image dataset [16, 120]. To design variable atlases dependent on subject information, a variety of approaches have been applied in recent years to the problem of characterizing anatomical changes in brain shape over time and during disease progression. Davis et al. [40] describe a method for population shape regression in which kernel regression is adapted to the manifold of diffeomorphisms and is used to obtain an age-dependent atlas. Ericsson et al. [48] propose a method for the construction of a patient-specific atlas where different average brain atlases are built in a small deformation setting according to meta-information such as sex, age, or clinical factors.

Methods for extracting features or biomarkers from MR brain image data often

begin by automatically segmenting regions of interest. A very popular segmentation technique is to use label propagation which transforms labels from an atlas image to an unseen target image by bringing both images into alignment. Atlases are typically, but not necessarily, manually labeled. Early work using this approach was proposed by Bajcsy et al. [11] as well as more recently Gee et al. [60] and Collins et al. [32]. The accuracy of label propagation strongly depends on the accuracy of the underlying image alignment. To overcome the reliance on a single segmentation, Warfield et al. [148] proposed STAPLE, a method that computes for a collection of segmentations a probabilistic estimate of the true segmentation. Rohlfing et al. [116] demonstrated the improved robustness and accuracy of a multi-classifier framework where the labels propagated from multiple atlases are combined in a decision-fusion step to obtain a final segmentation of the target image. Label propagation in combination with decision fusion was successfully used to segment a large number of structures in brain MR images by Heckemann et al. [73].

Due to the wide range of anatomical variation, the selection of atlases becomes an important issue in multi-atlas segmentation. The selection of suitable atlases for a given target helps to ensure that the atlas-target registrations and the subsequent segmentation are as accurate as possible. Wu et al. [155] describe different methods for improving segmentation results in the single atlas case by incorporating atlas selection. Aljabar et al. [1] investigate different similarity measures for optimal atlas selection during multi-atlas segmentation. Rikxoort et al. [143] propose a method where atlas combination is carried out separately in different sub-windows of an image until a convergence criterion is met. These approaches show that it is meaningful to select suitable atlases for each target image individually. Although an increasing number of MR brain images are available, the generation of high-quality manual atlases is a labor-intensive and expensive task (see e.g., [67]). This means that atlases are often relatively limited in number and, in most cases, restricted to a particular population (e.g. young, healthy subjects). This can limit the applicability of the atlas database even if a selection approach is used. To overcome this, Tang et al. [132] seek to produce

a variety of atlas images by utilizing a PCA model of deformations learned from transformations between a single template image and training images. Potential atlases are generated by transforming the initial template with a number of transformations sampled from the model. The assumption is that, by finding a suitable atlas for an unseen image, a fast and accurate registration to this template may be readily obtained. Test data with a greater level of variation than the training data would, however, represent a significant challenge to this approach. Additionally, the use of a highly variable training dataset may lead to an unrepresentative PCA model as the likelihood of registration errors between the diverse images and the single template is increased. This restriction makes this approach only applicable in cases were a good registration from all training images to the single initial template can be easily obtained.

The approach followed here aims to propagate a relatively small number of atlases through to a large and diverse set of MR brain images exhibiting a significant amount of anatomical variability. The initial atlases may only represent a specific subgroup of target image population and the method is designed to address this challenge. As previously shown, atlas-based segmentation benefits from the selection of atlases similar to the target image [155, 1]. Here, a framework is proposed that ensures this by first embedding all images in a low dimensional coordinate system that provides a distance metric between images and allows neighborhoods of images to be identified. In the manifold learned from coordinate system embedding, a propagation framework can be identified and labeled atlases can be propagated in a step-wise fashion, starting with the initial atlases, until the whole population is segmented. Each image is segmented using atlases that are within its neighborhood, meaning that deformations between dissimilar images are broken down to several small deformations between comparatively similar images and registration errors are reduced. To further minimize an accumulation of registration errors, an intensity-based refinement of the segmentation is done after each label propagation step. Once segmented, an image can in turn be used as an atlas in subsequent segmentation steps. After all images in the popula-

tion are segmented, they represent a large atlas database from which suitable subsets can be selected for the segmentation of unseen images. The coordinate system into which the images are embedded is obtained by applying a spectral analysis step [28] to their pairwise similarities. As labeled atlases are propagated and fused for a particular target image, the information they provide is combined with the intensity-model presented in Chapter 3.

The initial set of atlases used consists of the 30 atlases from young and healthy subjects described in Chapter 2.1.1. The proposed method is used to propagate this initial set of atlases to 796 ADNI baseline images 1.3. Results show that this approach provides more accurate segmentations due, at least in part, to the associated reductions in inter-subject registration error.

4.2 Materials and Methods

4.2.1 Subjects

The 796 available ADNI baseline images that were available in July 2009 were used for evaluation. An overview on the subjects is given in Table 4.1: For each subject group the number of subjects, the male/female distribution, the average age and the average result of the mini-mental stat examination (MMSE) [54] are shown.

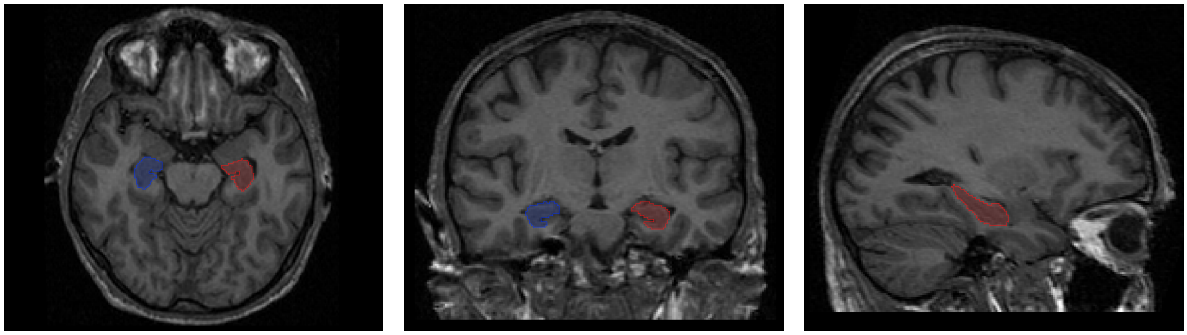
	N	M/F	Age	MMSE
Normal	222	116/106	76.00 ± 5.08 [60-90]	29.11 ± 0.99 [25-30]
MCI (all)	392	254/138	74.68 ± 7.39 [55-90]	27.02 ± 1.79 [23-30]
-S-MCI	230	155/75	74.88 ± 7.77 [55-90]	27.29 ± 1.80 [24-30]
-P-MCI	162	99/63	74.62 ± 6.96 [55-88]	26.63 ± 1.71 [23-30]
AD	182	91/91	75.84 ± 7.63 [55-91]	23.35 ± 2.00 [18-27]

Table 4.1: Information relating to the subjects whose images were used in this study.

For a subset of 182 of the 796 images, a semi-automated delineation for the hippocampus was provided by the ADNI consortium (Section A.1.1) and used as reference labels to evaluate the method.

4.2.2 Atlases

The initial set of atlases is defined by the 30 atlas images described in Chapter 2.1.1. Since no manual segmentations based on the Hammers protocol exists for the ADNI label maps used for evaluation of label overlaps, the definition of the hippocampus in the initial atlas was changed to make it consistent with manual hippocampus label maps provided by ADNI. An example of the ADNI delineation of the hippocampus on one of the 30 atlases is given in Figure 4.1.



(a) Transverse (b) Coronal (c) Sagittal

Figure 4.1: Manually delineated structures on a brain atlas

4.2.3 Overview of the method

To propagate an initial set of atlases through a dataset of images with a high level of inter-subject variance, a manifold representation of the dataset is learned where images within a local neighborhood are similar to each other. The manifold is represented by a coordinate embedding of all images. This embedding is obtained by applying a spectral analysis step [28] to the complete graph in which each vertex represents an image and all pairwise similarities between images are used to define the edge weights in the graph. Pairwise similarities can be measured as the intensity similarity between the images or the amount of deformation between the images or as a combination of the two.

In successive steps, atlases are propagated within the newly defined coordinate system. In the first step, the initial set of atlases are propagated to a number of images in their local neighborhood and used to label them. Images labeled in this way

become atlases themselves and are, in subsequent steps, further propagated throughout the whole dataset. In this way, each image is labeled using a number of atlases in its close vicinity which has the benefit of decreasing registration error. An overview on the segmentation process with the LEAP (Learning Embeddings for Atlas Propagation) framework is depicted in Figure 4.2.

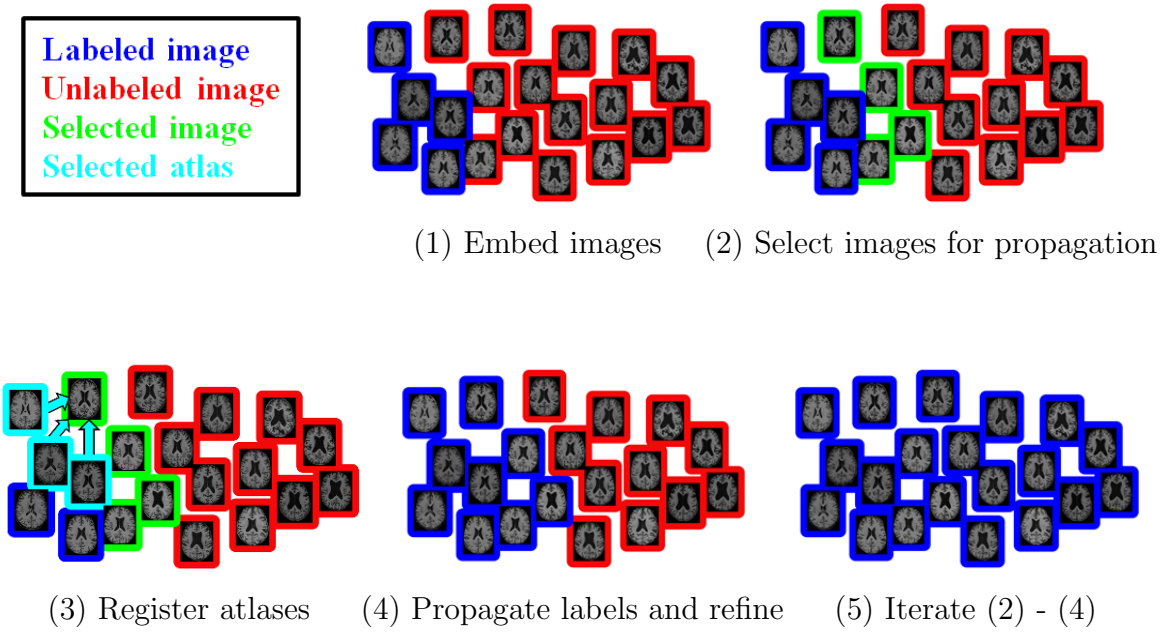


Figure 4.2: Process of atlas propagation with LEAP. All labeled (atlases) and unlabeled images are embedded into a low-dimensional manifold (1). The N closest unlabeled images to the labeled images are selected for segmentation (2). The M closest labeled images are registered to each of the selected images (an example for one image is shown in (3)). Intensity refinement is used to obtain label maps for each of the selected images (4). Steps (2) - (4) are iterated until all images are labeled.

4.2.4 Graph Construction and Manifold Embedding

In order to determine the intermediate atlas propagation steps, all images are embedded in a manifold represented by a coordinate system which is obtained by applying a spectral analysis step [28]. Spectral analytic techniques have the advantage of generating feature coordinates based on measures of pairwise similarity between data items such as images. This is in contrast to methods that require distance metrics between data items such as multidimensional scaling (MDS) (see Chapter 2.4). After a spectral analysis step, the distance between two images in the learned coordinate system is

dependent not only upon the original pairwise similarity between them but also upon all the pairwise similarities each image has with the remainder of the population. This makes the distances in the coordinate system embedding a more robust measure of proximity than individual pairwise measures of similarity which can be susceptible to noise. A good introduction to spectral analytic methods can be found in [145] and further details are available in [28].

The spectral analysis step is applied to the complete, weighted and undirected graph $G = (V, E)$ with each image in the dataset being represented by one vertex v_i. The non-negative weights w_{ij} between two vertices v_i and v_j are defined by the similarity s_{ij} between the respective images. In this work intensity based similarities are used (see Section 4.3.1). A weights matrix $\mathbf{W}$ for G is obtained by collecting the edge weights $w_{ij} = s_{ij}$ for every image pair and a diagonal matrix $\mathbf{T}$ contains the degree sums for each vertex $d_{ii} = \sum_j w_{ij}$. $\mathbf{T}$ gives a measure of how well every node is connected in the neighborhood graph. This reflects how similar an image is to the remainder of the population.

The normalized graph Laplacian $\mathcal{L}$ is then defined by [28]

$$\mathcal{L} = \mathbf{T}^{-1/2}(\mathbf{T} - \mathbf{W})\mathbf{T}^{-1/2}. \tag{4.1}$$

The Laplacian $\mathcal{L}$ encodes information relating to all pairwise relations between the vertices and the eigendecomposition of $\mathcal{L}$ provides a low-dimensional representation for each vertex[1]. The dimension of the low-dimensional space derived from a spectral analysis step can be chosen by the user. In this work, each dimension for the feature data was tested in turn while assessing the ability to discriminate between the four subject groups (young, AD, MCI and older control subjects). The discrimination ability was measured using the average inter-cluster distance based on the centroids of each cluster for each feature dimension. For the groups studied, it was maximal when using two-dimensional features and reduced thereafter (see Figure 4.3). A 2D

[1]The spectral embedding process described here is conceptually closely related to Laplacian eigenmaps as described in Chapter 2.4.2 [145]

representation is therefore used as a coordinate space in which to embed the data.

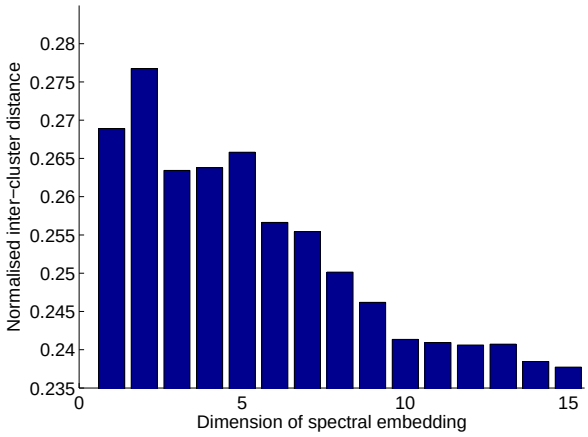


Figure 4.3: The discrimination ability for different chosen feature dimensions among the four subject groups (healthy young, elderly controls, MCI, AD). The best discrimination was achieved using a two dimensional embedding space which therefore was used to define the distances between images.

4.2.5 Segmentation Propagation in the Learned Manifold

In order to propagate the atlas segmentations through the dataset using the learned manifold, all images $I \in \mathbb{I}$ are separated into two groups, containing the *labeled* and *unlabeled* images. These groups are indexed by the sets $\mathbb{L}$ and $\mathbb{U}$ respectively. Initially, $\mathbb{L}$ represents the initial atlas images and $\mathbb{U}$ represents all other images. Let $d(I_i, I_j)$ represent the Euclidean distance between images I_i and I_j in the manifold, the average distance from an unlabeled image I_u to all labeled images is:

$$\bar{d}(I_u, \mathbb{L}) = \frac{1}{|\mathbb{L}|} \sum_{l \in \mathbb{L}} d(I_u, I_l) \tag{4.2}$$

At each iteration, the images I_u, $u \in \mathbb{U}$ with the N smallest average distances $\bar{d}(I_u)$ are chosen as targets for propagation. For each of these images, the M closest images drawn from I_l, $l \in \mathbb{L}$ are selected as atlases to be propagated. Subsequently, the index sets $\mathbb{U}$ and $\mathbb{L}$ are updated to indicate that the target images in the current iteration have been labeled. Stepwise propagation is performed in this way until all images in the dataset are labeled.

N is a crucial parameter as it determines the number of images labeled during each

iteration and therefore it strongly affects the expected number of intermediate steps that are taken before a target image is segmented. It needs to be set according to the diversity of the used dataset. A small value of N may be required in a very diverse dataset to guarantee that only 'similar' images need to be registered in every step. In a less diverse dataset, the value for N may be set to a larger value in order to avoid the unnecessary accumulation of registration errors. M defines the number of atlas images used for each application of multi-atlas segmentation. A natural choice is, to set M to the number of initial atlases. Independent of the choice of N, the number of registrations needed to segment K images is $M \times K$. The process of segmentation propagation in the learned manifold is summarized in Algorithm 1.

Algorithm 1 Segmentation propagation in the learned manifold

```
Set 𝕃 to represent the initial set of atlases
Set 𝕌 to represent all remaining images
while |𝕌| > 0 do
  for all I_u ∈ 𝕌 do
    calculate d̄(I_u, 𝕃)
  end for
  Reorder index set 𝕌 to match the order of d̄(I_u, 𝕃)
  for i = 1 to N do
    Select M images from I_l, l ∈ 𝕃 that are closest to I_{u_i}
    Register the selected atlases to I_{u_i}
    generate a multi-atlas segmentation estimate of I_{u_i}
  end for
  Transfer the indices {u_1, ..., u_N} from 𝕌 to 𝕃
end while
```

4.2.6 Multi-atlas propagation and segmentation refinement

Each label propagation is carried out by applying the graph-cuts based method described in Chapter 3. By incorporating intensity information from the unseen image into the segmentation process, errors done with conventional multi-atlas segmentation can be overcome [140, 153].

Each registration used to build the subject-specific probabilistic atlas (see Chapter 3, Equation 3.2) is carried out in three steps: rigid, affine and non-rigid. Rigid and affine registrations are carried out to correct for global differences between the images.

In the third step, two images are non-rigidly aligned using a free-form deformation model in which a regular lattice of control point vectors are weighted using B-spline basis functions to provide displacements at each location in the image [118]. The deformation is driven by the normalized mutual information [130] of the pair of images. The spacing of B-spline control points defines the local flexibility of the non-rigid registration. A sequence of control point spacings was used in a multi-resolution fashion (20mm, 10mm, 5mm and 2.5mm).

4.3 Experiments and Results

4.3.1 Image similarities

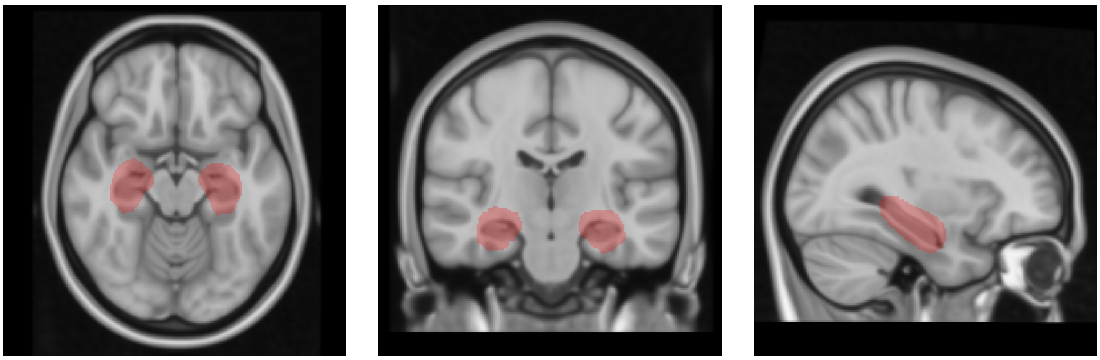
An intensity-based similarity between a pair of images I_i and I_j is used in this application. This similarity is based on normalized mutual information (NMI) [130] which is with the entropy $H(I)$ of an image I and the joint entropy $H(I_i, I_j)$ of two images defined as

$$NMI_{ij} = \frac{H(I_i) + H(I_j)}{H(I_i, I_j)} \tag{4.3}$$

For the first part of the evaluation, that aims at accurately segmenting the hippocampus, the similarity measure between a pair of images is estimated as the NMI over a region of interest (ROI) around the hippocampus. In the second part of the evaluation, the influence of using whole-brain similarities in contrast to a ROI is assessed. The framework is general and a user can choose the similarity measure and region of interest appropriate to the region or structure being segmented.

To define the hippocampus ROI, all training images were automatically segmented using standard multi-atlas segmentation [73]. The resulting hippocampal labels were then aligned to the MNI152-brain T1 atlas [106] using a coarse non-rigid registration modeled by free-form deformations (FFDs) with a 10mm B-spline control point spacing [118] between the corresponding image and the atlas. The hippocampal ROI was then

defined through the dilation of the region defined by all voxels which were labeled as hippocampus by at least 2% of the segmentations. To evaluate the pairwise similarities, all images were aligned to the MNI152-brain atlas using the same registrations used for the mask building. Figure 4.4 shows the ROI around the hippocampus superimposed on the brain atlas used for image normalization.



(a) Transverse (b) Coronal (c) Sagittal

Figure 4.4: The MNI152 brain atlas showing the region of interest around the hippocampus that was used for the evaluation of pairwise image similarities

4.3.2 Coordinate system embedding

The method for coordinate system embedding described in Section 4.2.4 was applied to a set of images containing the 30 initial atlases and the 796 ADNI images. The first two features from spectral graph analysis were used to embed all images into a 2D coordinate system. The results of coordinate system embedding are displayed in Figure 4.5. The original atlases form a distinct cluster on the left hand side of the graph at low values for the first feature. Furthermore it can be seen that control subjects are mainly positioned at lower values, whereas the majority of AD subjects is positioned at higher values. The hippocampal area for chosen example subjects is displayed in Figure 4.5. These types of observations support the impression that neighborhoods in the coordinate system embedding represent images that are similar in terms of hippocampal appearance.

All 796 images were segmented using five different approaches:

I Direct segmentation using standard multi-atlas segmentation [73].

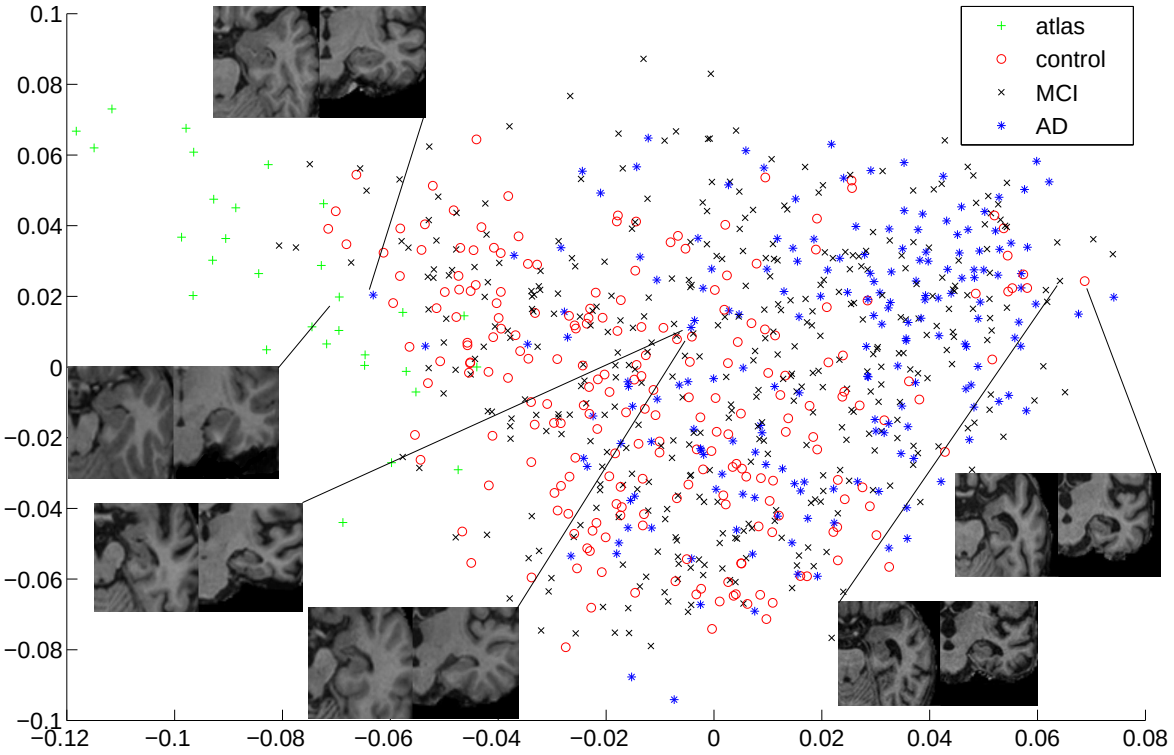


Figure 4.5: Abscissa and ordinate show first and second coordinates respectively of a low-dimensional embedding space. Embedded are 30 atlases based on healthy subjects and 796 images from elderly dementia patients and age matched control subjects. Details of images showing the hippocampus in example subjects.

II Direct segmentation using multi-atlas segmentation in combination with an intensity refinement based on graph cuts [140, 153] (see Chapter 3).

III LEAP with M=30 and N=300 and no intensity refinement after multi-atlas segmentation.

IV LEAP (see Section 4.2.2) with M=30 and N=1.

V LEAP with M=30 and N=300.

4.3.3 Evaluation of hippocampus segmentations

For evaluation, the automatic segmentation of the ADNI images were compared with the semi-automated and manually corrected hippocampus segmentations described in Appendix A.1.1. This comparison was carried out for all of the images for which ADNI provides a reference segmentation (182 out of 796). Comparing these 182 subjects (Table 4.2) with the entire population of 796 subjects (Table 4.1) shows that the

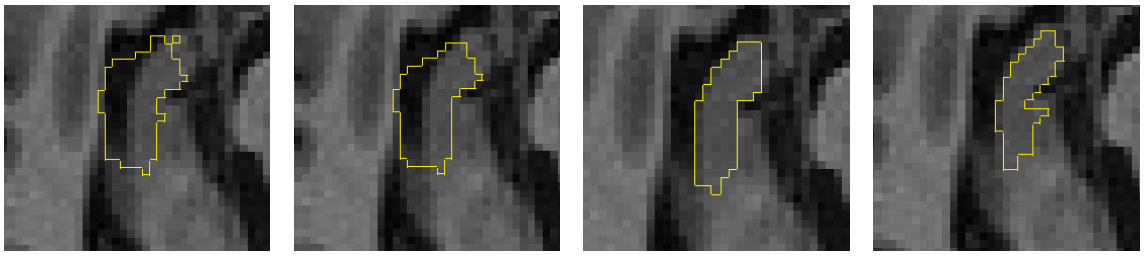
subgroup is characteristic of the entire population in terms of age, sex, MMSE and pathology.

	N	M/F	Age	MMSE
Normal	57	27/30	77.1 ± 4.60 [70-89]	29.29 ± 0.76 [26-30]
MCI	84	66/18	76.05 ± 6.77 [60-89]	27.29 ± 3.22 [24-30]
AD	41	21/20	76.08± 12.80 [57-88]	23.12 ± 1.79 [20-26]

Table 4.2: Characteristics of the subjects used for comparison between manual and automatic segmentation

An example for the segmentation of the right hippocampus of an AD subject is shown in Figure 4.3.3. A clear over-segmentation into CSF space and especially an under-segmentation in the anterior part of the hippocampus can be observed, both in the case of multi-atlas segmentation with and without intensity-based refinement (methods I and II). The fact that the intensity-based refinement cannot compensate for this error is due to the high spatial prior in this area that is caused by a significant misalignment of the majority of atlases in this area. The resulting high spatial prior cannot be overcome by the intensity-based correction scheme. When using the proposed framework without intensity-refinement (method III), the topological errors can be avoided, but the over-segmentation into CSF space is still present. The figure also shows that all observed problems can be avoided by using the proposed framework.

The average overlaps as measured by the Dice coefficient or similarity index (SI) [42] for the segmentation of left and right hippocampus on the 182 images used for evaluation are shown in Table 4.3. The difference between all pairs of the five methods is statistically significant with $p < 0.001$ on Student's two-tailed paired t-test.



(a) direct (b) direct, GC (c) LEAP, N=300 (d) manual

Figure 4.6: Comparison of segmentation results for the right hippocampus on a transverse slice.

	left hippocampus	right hippocampus
direct	0.775 ± 0.087 [0.470-0.904]	0.790 ± 0.080 [0.440-0.900]
direct, GC	0.820 ± 0.064 [0.461-0.903]	0.825 ± 0.065 [0.477-0.901]
LEAP, N=300, no GC	0.808 ± 0.054 [0.626-0.904]	0.814 ± 0.053 [0.626-0.900]
LEAP,N=1	0.838 ± 0.023 [0.774-0.888]	0.830 ± 0.024 [0.753-0.882]
LEAP,N=300	0.848 ± 0.033 [0.676-0.903]	0.848 ± 0.030 [0.729-0.905]

Table 4.3: Similarity index (SI) for hippocampus segmentation.

These results clearly show an improved segmentation accuracy and robustness for the proposed method. A hypothesis is that by avoiding the direct registration of images whose distance in the embedded space is too large but instead registering the images via multiple intermediate images improves significantly the segmentation accuracy and robustness of multi-atlas segmentation. To test this hypothesis, the development of segmentation accuracy was evaluated as a function of distances in the coordinate system embedding as well as the number of intermediate steps. Figure 4.7 shows this for the five segmentation methods in the form of ten bar plots: Each bar plot corresponds to the average SI overlap of 18 images (20 in the last plot). The first plot represents the 18 images closest to the original atlases, the next plot represents images slightly further from the original atlases and so on. These results show the superiority of the proposed method over direct multi-atlas segmentation approaches in segmenting images that are different from the original atlas set.

With increasing distance from the original atlases in the learned manifold, the accuracy of direct multi-atlas segmentation (method I) as well as multi-atlas segmentation with intensity-based refinement (method II) steadily decreases. By contrast, LEAP with both parameter settings shows a steady level of segmentation accuracy. It is interesting to see, that the described method with a step width of $N = 1$ (method IV) leads to worse results than the direct multi-atlas methods up to a certain distance from the original atlases. This can be explained by registration errors accumulated through many registration steps. With increasing distance from the atlases, however, the gain from using intermediate templates, outweighs this registration error. Furthermore, the accumulated registration errors do not seem to increase dramatically after a certain number of registrations. This is partly due to the intensity-based correction in every

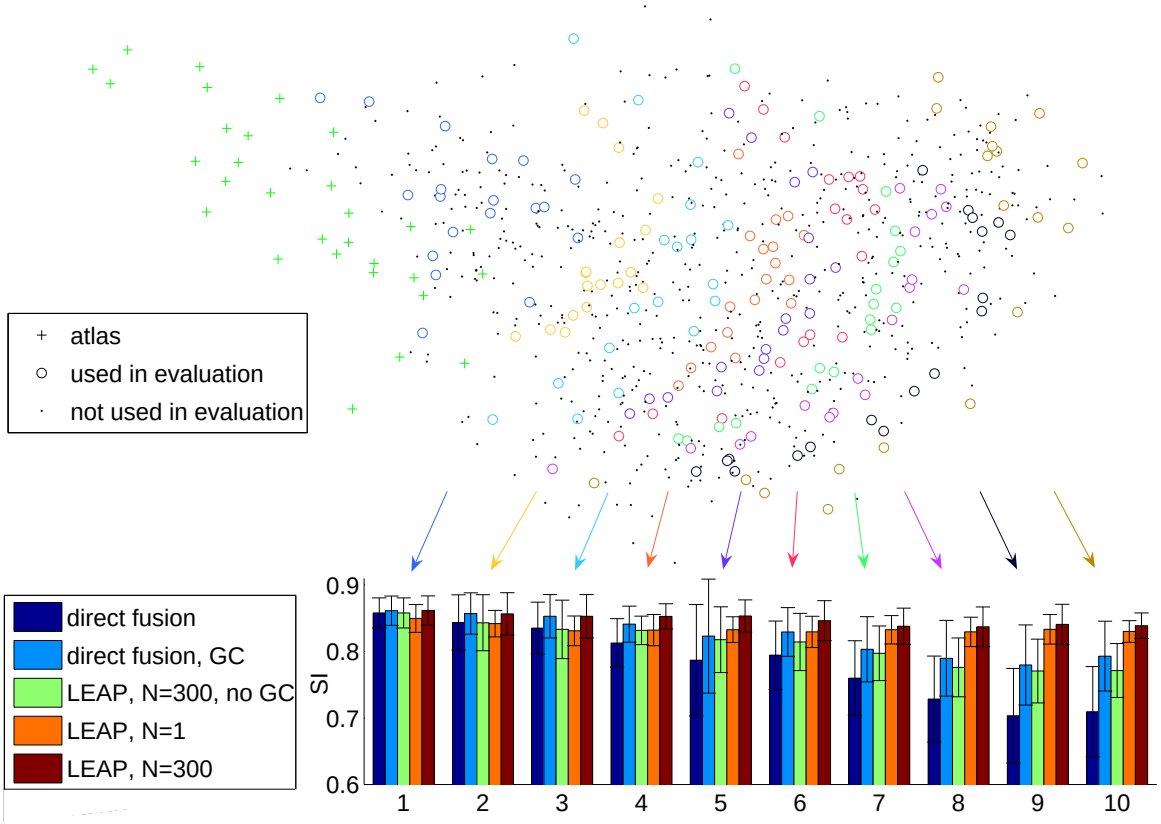
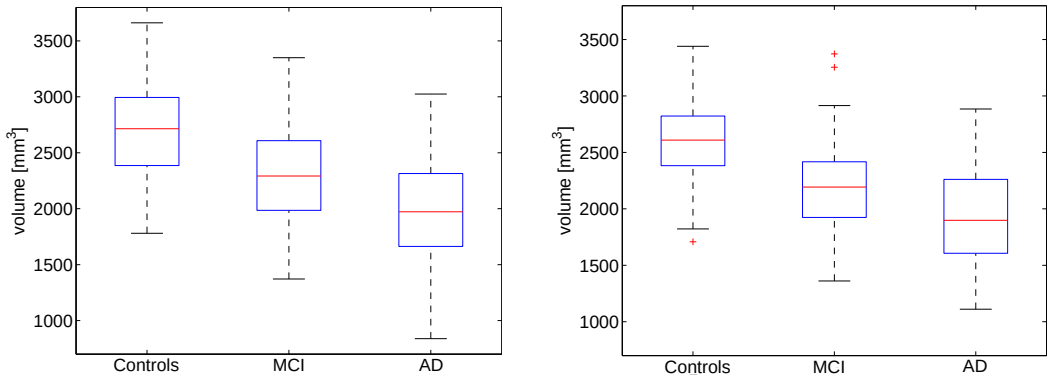


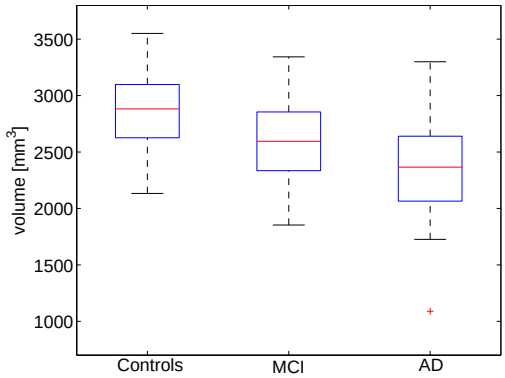
Figure 4.7: Development of segmentation accuracy with increasing distance from the original set of atlases. Each subset of images used for evaluation is represented by one bar plot.

multi-atlas segmentation step which corrects for small registration errors. Segmenting the 300 closest images with LEAP before doing the next intermediate step ($N = 300$, method V), leads to results at least as good as and often better than those given by the direct methods for images at all distances from the initial atlases. The importance of an intensity-based refinement step after multi-atlas segmentation is also underlined by the results of method III. When applying LEAP without this step, the gain compared to method I gets more and more significant with more intermediate steps, but the accuracy still declines significantly which can be explained by a deterioration of the propagated atlases (note that for the first 300 images, method II and method V are identical, as are methods I and III). The influence of N on the segmentation accuracy is governed by the trade-off between using atlases that are as close as possible to the target image (small N) and using a design where a minimum number of intermediate steps are used to avoid the accumulation of registration errors (large N). Due to the

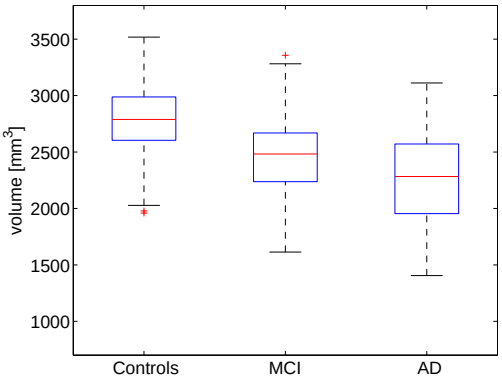


(a) manual, right hippocampus

(b) manual, left hippocampus



(c) automatic, right hippocampus



(d) automatic left hippocampus

Figure 4.8: Average hippocampal volumes for manual and automatic segmentation using method IV.

computational complexity of evaluating the framework, the evaluation was restricted to two values for N.

4.3.4 Volume measurements

A reduction in hippocampal volume is a well-known factor associated with cognitive impairment (e.g. [80, 115]). To measure the ability of our method to discriminate clinical groups by hippocampal volume, we compared the volumes measured on the 182 manually labeled images to the ones obtained from our automatic method (method V, LEAP with $M = 30$ and $N = 300$). Boxplots showing these volumes for the left and right hippocampus are displayed in Figure 4.8. The discriminative power for the volume of left and right hippocampus between all pairs of clinical groups is statistically significant with $p < 0.05$ on a Student's t-test but is slightly less significant than the manual discrimination. The power of automatically derived volumes to discriminate

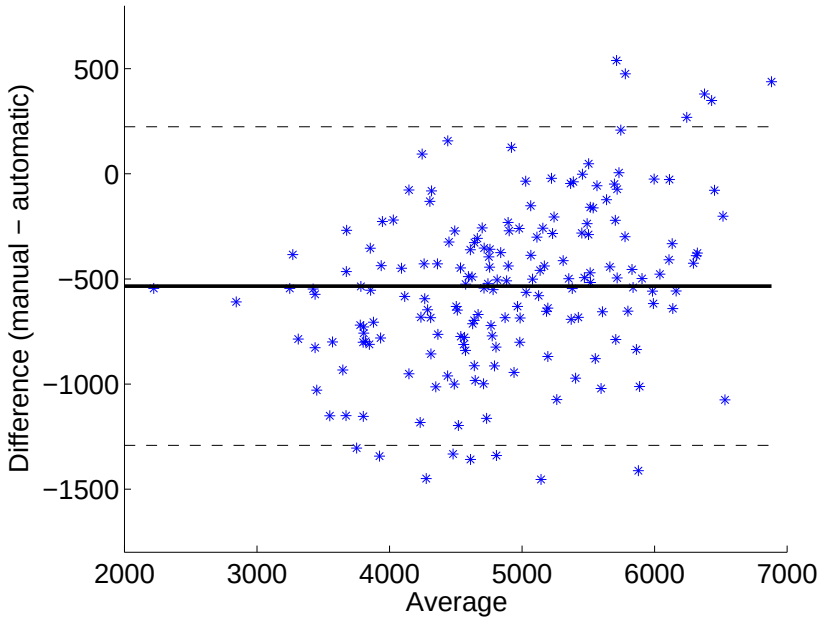


Figure 4.9: A Bland-Altman plot showing the agreement between volume measurement based on manual- and automatic segmentation of the hippocampus (method IV). The solid line represents the mean and the dashed lines represent ±1.96 standard deviations.

between clinical subject groups is presented in Section 4.3.5.

A Bland-Altman plot of the agreement of the two volume measurements is shown in Figure 4.9. This plot supports the impression of the volume measures in Figure 4.8 that the automated method tends to slightly overestimate the hippocampal volumes. This over-segmentation is more significant for small hippocampi. The same phenomenon has been described for an automatic segmentation method before by [68]. The intraclass correlation coefficient (ICC) between the volume measurements based on the manually corrected and automatic segmentation is 0.898 (ICC (3,1) Shrout-Fleiss reliability [125]). This value is comparable to the value of 0.929 reported in [112] for inter-rater reliability.

4.3.5 Segmentation of 83 brain structures

To further evaluate the proposed LEAP framework, its application to all 83 anatomical structures in the used atlas is evaluated. Since no manual labels based on the atlas protocol are available for the ADNI data, classification accuracy between clinical groups is evaluated to test the ability of the derived structural volumes to serve as a biomarker for AD.

Segmentation using whole brain similarities

The segmentation propagation of hippocampal label maps with LEAP presented in Section 4.3.3 is based on pairwise similarities evaluated over an ROI around the hippocampus. For a global distance measure, a manifold embedding for the whole brain is used here to propagate the whole brain atlas. To accommodate for inter-subject differences on a coarser level, an affine registration is used to measure pairwise whole brain similarities in contrast to the non-rigid registration described in Section 4.3.1. After registering all subjects to the MNI152 brain template, pairwise similarities are evaluated over the whole brain as described in Equation 4.3. The first two embedding coordinates from applying the spectral embedding step to the obtained whole brain similarity matrix is presented in Figure 4.10.

Comparing this embedding to the embedding based on hippocampal appearance and presented in Figure 4.5, it can be seen that the manually labeled young and healthy atlas subjects are still clustered on the left hand side. However, the discrimination between healthy controls and AD subjects is less clear in the embedding based on whole brain similarities. This result is expected when considering the more significant changes in hippocampal appearance related to the development of AD.

Based on inter-subject distances in the manifold based on whole brain similarities, LEAP is applied to all 83 defined atlas regions. Intensity-based refinement as proposed in Chapter 3 is applied to structures with a homogeneous gray-value only. These are structures 1-4, 19, 34-46, 49, 74 and 75 from the list given in Table B.1.

Extracted volumes for all 83 regions are used as a feature to discriminate between

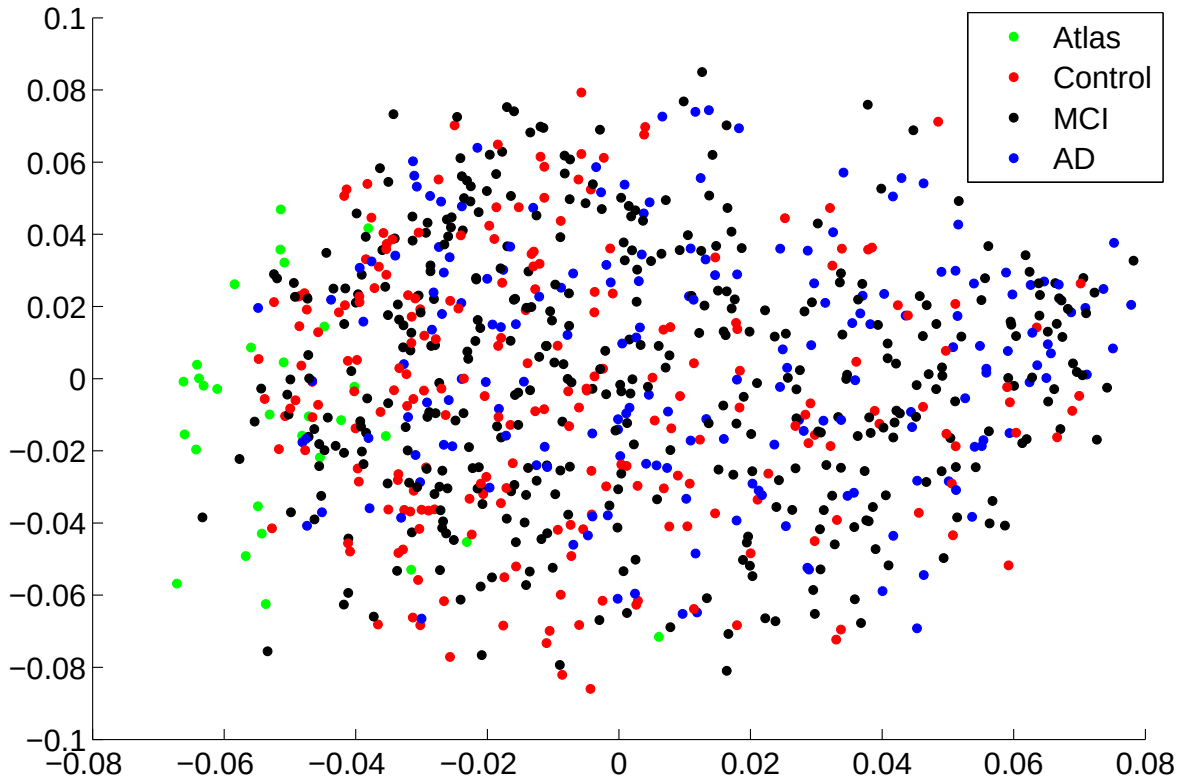


Figure 4.10: First two embedding coordinate for 796 ADNI images together with 30 manually labeled atlas images obtained from applying a spectral embedding step to pairwise similarities evaluated over the whole brain after an affine normalization to a template space.

clinically relevant subject groups. The volumes used for classification in all experiments described in this chapter are corrected for subject age using a multiple linear regression model. Support vector machines (SVM's) are used in a leave-25%-out fashion to discriminate AD subjects from healthy controls (AD vs CN) as well as progressive MCI subjects from healthy controls (P-MCI vs CN) and stable MCI subjects (P-MCI vs S-MCI). Classification accuracies for the three comparisons with automatically extracted volumes for all 83 structures are displayed in Figure 4.11. Classification accuracy (ACC), sensitivity (SEN) and specificity (SPE) for the volumes that at least for one comparison performed better than chance (sensitivity and specificity larger than zero), are presented in Table 4.4.

To explore the potential of several volume measurements to improve classification accuracy, different methods to compare more than one measure were compared. In the first method, SVM-based classification was applied in the d-dimensional space formed

Structure	AD vs CN ACC / SEN / SPE	S-MCI vs P-MCI ACC / SEN / SPE	CN vs P-MCI ACC / SEN / SPE
Hippocampus, r	72.6 / 64.4 / 79.4	59.3 / 18.4 / 87.2	70.3 / 56.6 / 80.0
Hippocampus, l	74.1 / 70.1 / 77.4	**62.1** / 31.1 / 83.2	72.4 / 63.9 / 78.5
Amygdala, r	**77.6** / 72.1 / 82.2	61.5 / 31.6 / 81.9	**72.7** / 59.7 / 81.9
Amygdala, l	74.5 / 69.9 / 78.2	63.0 / 33.7 / 82.9	70.0 / 57.9 / 78.6
Gyri parahippocampalis, r	65.6 / 46.2 / 81.5	59.0 / 0.1 / 99.1	59.4 / 11.9 / 93.2
Gyri parahippocampalis, l	63.4 / 47.5 / 76.5	59.9 / 5.4 / 97.0	63.1 / 26.5 / 89.1
Sup. temporal gyru (post.), r	56.3 / 28.8 / 78.9	59.3 / 1.3 / 98.9	57.1 / 5.3 / 93.9
Sup. temporal gyru (post.), l	57.2 / 27.9 / 81.3	59.4 / 4.1 / 97.0	57.7 / 2.8 / 96.8
Med. and inf. temp. gyri, r	60.6 / 32.5 / 83.6	60.6 / 7.3 / 96.9	56.8 / 2.9 / 95.2
Med. and inf. temp. gyri, l	59.2 / 33.4 / 80.3	60.1 / 11.4 / 93.3	61.2 / 22.0 / 89.1
Cerebellum, l	56.1 / 20.7 / 85.1	59.5 / 0.0 / 100.0	58.4 / 0.0 / 99.9
Brainstem	57.8 / 28.1 / 82.3	59.5 / 0.0 / 100.0	58.9 / 14.4 / 90.5
Insula, l	56.0 / 24.7 / 81.7	59.0 / 1.6 / 98.1	58.6 / 6.1 / 95.9
Occipital lobe, r	57.7 / 14.2 / 93.3	59.3 / 2.7 / 97.8	58.2 / 1.6 / 98.5
Cingulate gyrus (anterior), l	58.0 / 27.7 / 82.7	58.9 / 0.3 / 98.9	58.4 / 1.5 / 99.0
Cingulate gyrus (anterior), r	56.9 / 24.1 / 83.9	58.8 / 0.0 / 98.9	57.6 / 3.7 / 95.9
Posterior temporal lobe, l	54.7 / 17.2 / 85.5	60.4 / 6.5 / 97.0	58.0 / 3.1 / 97.1
Posterior temporal lobe, r	57.2 / 22.6 / 85.5	58.9 / 4.0 / 96.3	57.9 / 3.4 / 96.7
Nucleus accumbens, l	60.1 / 44.5 / 72.9	59.3 / 4.2 / 96.9	61.1 / 24.6 / 87.1
Nucleus accumbens, r	55.8 / 2.3 / 99.6	59.5 / 2.8 / 98.1	60.7 / 14.4 / 93.6
Putamen, r	57.8 / 23.2 / 86.2	58.8 / 0.9 / 98.1	59.6 / 23.2 / 85.5
Thalamus, l	64.0 / 49.7 / 75.7	59.2 / 3.7 / 97.0	64.7 / 40.1 / 82.2
Thalamus, r	62.9 / 47.7 / 75.4	59.1 / 2.3 / 97.8	66.2 / 42.3 / 83.2
Pallidum, globus pallidus, l	57.4 / 22.2 / 86.2	59.4 / 0.0 / 99.9	57.6 / 2.6 / 96.8
Corpus callosum	59.5 / 38.7 / 76.5	60.0 / 3.5 / 98.4	59.2 / 22.4 / 85.3
Lat. ventricle, front. horn., r	66.3 / 46.6 / 82.5	57.7 / 3.3 / 94.7	63.5 / 32.0 / 85.9
Lat. ventricle, front. horn., l	62.3 / 40.2 / 80.4	57.7 / 1.9 / 95.7	61.0 / 30.7 / 82.6
Lat. ventricle, temp. horn, r	68.4 / 45.5 / 87.1	60.7 / 16.6 / 90.7	65.3 / 33.1 / 88.3
Lat. ventricle, temp. horn, l	69.7 / 43.5 / 91.2	60.5 / 12.0 / 93.6	65.6 / 28.3 / 92.2
Third Ventricle	60.3 / 39.2 / 77.6	58.8 / 1.9 / 97.6	58.4 / 16.0 / 88.6
Sup. parietal gyrus, l	57.7 / 18.5 / 89.9	59.0 / 0.1 / 99.1	58.2 / 1.2 / 98.7
Sup. parietal gyrus, r	57.2 / 18.1 / 89.3	58.8 / 0.4 / 98.6	57.6 / 0.6 / 98.2
Medial orbital gyrus, l	55.9 / 9.3 / 94.1	59.8 / 7.1 / 95.7	59.7 / 13.0 / 93.0
Medial orbital gyrus, r	56.9 / 14.2 / 91.9	58.9 / 2.3 / 97.5	59.5 / 13.7 / 92.1

Table 4.4: Classification accuracy (ACC), sensitivity (SEN) and specificity (SPE) achieved with support vector machines (SVMs) based on automatically determined volumes. Results are displayed for all structures for which the classification accuracy of at least one comparison performed better than chance.

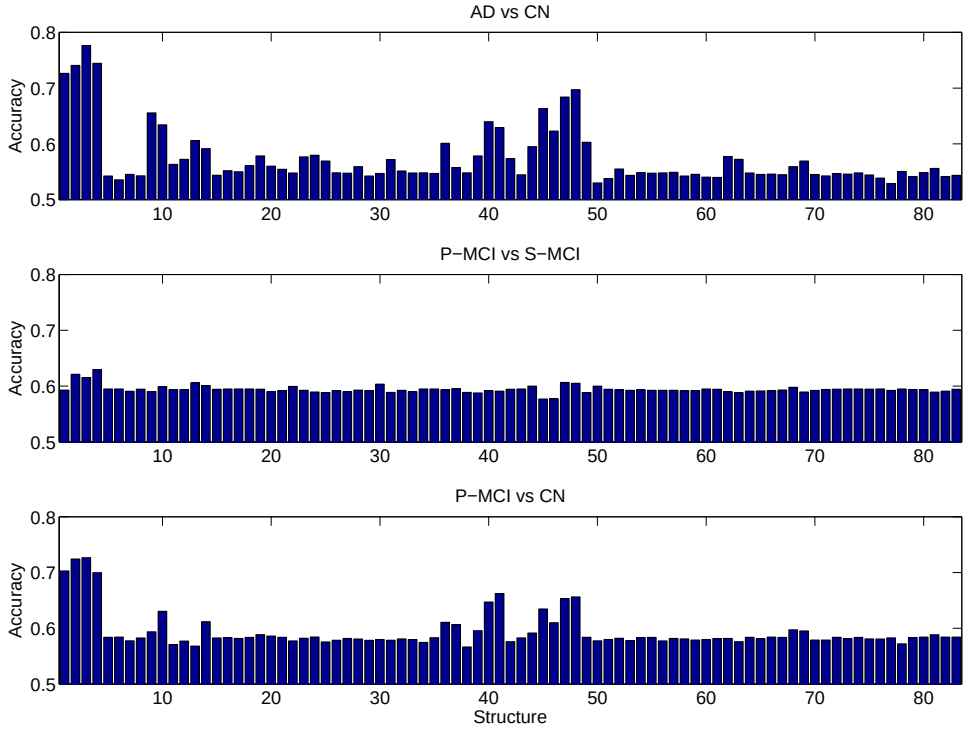


Figure 4.11: Classification accuracy for three different clinical groupings achieved from 83 delineated brain structures.

by d independent volumetric measures. In the second method, AdaBoost [59] is used to find a suitable classifier from a set of volumes. In AdaBoost, a strong classifier is defined by iteratively selecting weak classifiers to improve it's performance on a training dataset. Similar to the SVM-based approach, AdaBoost was applied on the volumes by applying a leave-25%-out approach.

Table 4.4 shows that the best performance for every classification task is achieved with individual volumes obtained from either hippocampus or amygdala. The four volumes from these two structures are therefore used to define a first feature set (set I). The second set (set II) is defined by all structures represented in Table 4.4 with which a discrimination between at least one pairing of clinical groups with an accuracy higher than chance is possible. The third set (set III) consists of all 83 structural volumes.

Classification accuracy, sensitivity and specificity for SVM and AdaBoost with volume sets I, II and III are displayed in Table 4.5. Combining different volumes shows a

clear improvement in classification accuracy, in particular for the discrimination from healthy control subjects. Best classification rates are achieved when applying SVMs to feature set II and AdaBoost to all 83 structures (set III). However, AdaBoost shows a significantly better performance on the S-MCI vs P-MCI classification. Furthermore, feature set II is build based on classification performance which makes it less suitable for an application in practice. It can also be seen that the application of the boosting approach to the relatively small set I (4 volumes) performs worse than a direct application of SVM. This could be explained by the hypothesis that classifier selection only helps to improve accuracy if a large number of weak classifiers are available.

Regional similarities

In this section, the influence of using a local similarity measure as done in Section 4.3.1 in contrast to using a global measure as done in Section 4.3.5 is assessed. In a first test, the overlaps of an automatically derived hippocampus segmentation using the two measures with 182 reference hippocampus segmentations (Appendix A.1.1) are compared. Figure 4.12 shows average SI overlaps for the 10 subject groupings used in Figure 4.7 when using whole brain similarities (blue) and hippocampus similarities (grey) to define LEAP propagation with N=300 labelings per step and M=20 atlases. Average SI values for all 182 subjects are SI=0.845±0.032 for whole brain similarities and SI=0.848±0.027 for hippocampus similarities. While the results obtained with

Method / volume set	AD vs CN ACC / SEN / SPE	S-MCI vs P-MCI ACC / SEN / SPE	CN vs P-MCI ACC / SEN / SPE
SVM			
I: Hippo./Amygdala	78.7 / 74.2 / 82.4	61.5 / 35.3 / 79.4	75.2 / 67.3 / 80.7
II: selected	83.8 / 79.9 / 86.9	58.7 / 40.1 / 71.3	77.6 / 70.2 / 82.9
III: all 83 structures	82.3 / 78.9 / 85.1	59.4 / 44.7 / 69.4	76.8 / 70.1 / 81.5
AdaBoost			
I: Hippo./Amygdala	72.9 / 68.3 / 76.6	59.5 / 48.7 / 66.8	68.8 / 59.8 / 75.2
II: selected	83.0 / 78.8 / 86.4	60.9 / 48.2 / 69.5	75.8 / 67.3 / 81.8
III: all 83 structures	83.2 / 78.6 / 86.9	63.2 / 49.3 / 72.6	76.3 / 66.9 / 82.9

Table 4.5: Classification results achieved with the combination of different structural volumes. Support vector machines (SVM) and AdaBoost were used on different sets of volumes to perform classification.

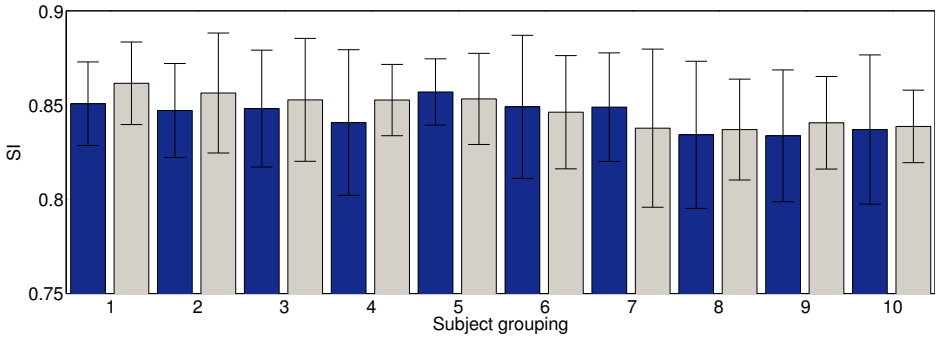


Figure 4.12: Label overlaps (SI) for automated hippocampus segmentation with semi-automated reference segmentations. Compared are segmentations obtained with LEAP when using a similarity measure over the whole brain (blue) to a similarity measure defined in a region around the hippocampus (grey). Results for both approaches are represented for 10 groups of subjects as described before for Figure 4.7.

hippocampus similarities show slightly higher mean values with a lower standard deviation, a paired t-test shows no significant difference between the two distributions with $p = 0.32$.

In a second test, the influence of the input similarity to LEAP on the classification accuracy with the resulting volumes is examined on three exemplar structures. Apart from the hippocampus, LEAP was independently applied to amygdala and the parahippocampal gyrus, where the latter is an example for a structure for which, due to it's inhomogeneity, no intensity-based correction after every atlas-propagation step is applied. Obtained classification accuracies after measuring combined volumes (right + left) based on local similarities, are displayed in the bottom part of Table 4.6. In addition to the rates achieved with individual volumes, the combination of all volumes is used for SVM- and AdaBoost-based classification as described in the previous Section. Classification results are at least as good as the ones for a global similarity measure presented in Table 4.4. For comparison, the top part of Table 4.6 shows the classification accuracy achieved with the volumes obtained from directly registering the manually labeled atlases to all target images and performing intensity-based refinement, where applicable.

Method / Structure	AD vs CN ACC / SEN / SPE	S-MCI vs P-MCI ACC / SEN / SPE	CN vs P-MCI ACC / SEN / SPE
Direct propagation			
Hippocampus	68.6 / 50.3 / 83.6	60.5 / 10.3 / 94.6	67.1 / 50.9 / 78.7
Amygdala	73.7 / 64.0 / 81.7	62.7 / 34.6 / 81.8	74.0 / 64.4 / 80.7
Gyri parahippocampalis	59.3 / 28.4 / 84.6	59.0 / 0.0 / 100.0	59.0 / 10.0 / 93.9
SVM combined	77.4 / 73.1 / 81.0	61.9 / 33.7 / 81.2	73.7 / 62.9 / 81.3
Adaboost combined	69.6 / 64.5 / 74.5	55.8 / 40.7 / 66.0	63.3 / 52.2 / 71.2
LEAP Local			
Hippocampus	77.0 / 71.0 / 81.9	61.0 / 27.9 / 83.5	74.8 / 64.3 / 82.3
Amygdala	79.8 / 75.5 / 83.3	62.7 / 36.2 / 80.8	76.2 / 67.7 / 82.3
Gyri parahippocampalis	72.7 / 63.7 / 80.0	60.6 / 15.4 / 91.4	70.5 / 50.9 / 84.4
SVM combined	79.8 / 73.6 / 84.1	63.8 / 35.7 / 82.9	75.2 / 62.3 / 84.4
Adaboost combined	73.3 / 69.3 / 76.6	61.1 / 47.9 / 70.2	67.1 / 56.8 / 74.4

Table 4.6: Classification accuracies achieved for selected structural volumes. The top part of the table shows results after direct propagation of atlas images, the bottom part shows the results after applying LEAP where pairwise similarity is measured in a region around the structure of interest.

4.4 Discussion and Conclusion

This chapter described the LEAP framework for propagating an initial set of brain atlases to a diverse population of unseen images via multi-atlas segmentation. The process starts by embedding all atlas and target images in a coordinate system where similar images according to a chosen measure are close. The initial set of atlases is then propagated in several steps through the manifold represented by this coordinate system. This avoids the need to estimate large deformations between images with significantly different anatomy and the correspondence between them is broken down into a sequence of comparatively small deformations. The formulation of the framework is general and is not tied to a particular similarity measure, coordinate embedding or registration algorithm.

LEAP was applied to a dataset of 796 images acquired from elderly dementia patients and age matched controls using a set of 30 atlases of healthy young subjects. In a first step, the method was applied to the task of hippocampal segmentation and consistently improved segmentation results were achieved compared to standard multi-atlas segmentation. Furthermore, a consistent level of accuracy for the proposed approach was achieved with increasing distance from the initial set of atlases and therefore with

more intermediate registration steps. The accuracy of standard multi-atlas segmentation, on the other hand, steadily decreased. This observation suggests three main conclusions: 1) The decreasing accuracy of the standard multi-atlas segmentation suggests that the coordinate system embedding used is meaningful. The initial atlases get less and less suitable for segmentation with increasing distance. 2) The almost constant accuracy of the proposed method suggests that, by using several small deformations, it is possible to indirectly deform an atlas appropriately to a target in a way that is not matched by a direct deformation with multi-atlas segmentation. 3) The gain from restricting registrations to similar images counters the accumulation of errors when using successive small deformations.

The presented results indicate that, if many intermediate registrations are used, the segmentation accuracy initially declines quickly but then remains relatively constant with increasing distance from the initial atlases. The initial decline can be explained by an accumulation of registration errors which results from many intermediate registration steps. The reason why the accuracy does not monotonically decline is likely to be due to the incorporation of the intensity model during each multi-atlas segmentation step. By automatically correcting the propagated segmentation based on the image intensities, the quality of the atlas can be preserved to a certain level.

In further tests, the presented framework was applied to whole-brain segmentation, evaluating 83 structural volumes in 796 images. Using classification accuracy as an indicator of the quality of the segmentation, the ability of automatically determined volumes to classify between different clinical subject groups was tested. The results show that combining multiple volumes can substantially improve classification rates. Furthermore, based on three exemplar regions, it could be observed that structural volumes obtained with LEAP perform significantly better in discriminating clinical groups than volumes obtained from standard multi-atlas segmentation (Table 4.6).

Further tests evaluated the sensitivity of the LEAP framework to the similarity measure used to define the low-dimensional manifold space which in turn is used to determine the step-wise propagation scheme. When using hippocampal label over-

lap as a measure of accuracy, no significant difference can be observed between using whole-brain similarities and using similarities specifically evaluated in a region around the hippocampus. Two main conclusions can be drawn from this finding: 1) there is a correlation between similarities evaluated over the whole brain and similarities evaluated over the hippocampal ROI. This is plausible when assuming that whole brain similarities are mainly influenced by ventricular appearance and a correlation between ventricular- and hippocampal atrophy. 2) The proposed framework is relatively robust with respect to the segmentation order in the learned manifold. This can be explained by the registration algorithm being able to 'bridge the gap' between different images as long as they are 'reasonable' similar. In addition, mis-labellings occurred by registration are, up to a certain level, corrected by the intensity-based refinement step.

Apart from the obvious application of segmenting a dataset of diverse images with a set of atlases based on a sub-population, the proposed method can be seen as an automatic method for generating a large repository of atlases for subsequent multi-atlas segmentation with atlas selection [1]. Since the manual generation of large atlas databases is expensive, time-consuming and in many cases unfeasible, the proposed method could potentially be used to automatically generate such a database.

Notwithstanding the challenge represented by variability due to image acquisition protocols and inter-subject variability in a dataset as large and as diverse as the one in the ADNI-study, the results achieved with the proposed method compare well to state of the art methods applied to more restricted datasets [140, 108, 30, 68] in terms of accuracy and robustness.

The next chapter presents an extension of this framework to longitudinal image datasets so that atrophy rates can be accurately determined.

Chapter 5

Consistent segmentation of longitudinal images to measure atrophy

This chapter is based on:

Robin Wolz, Rolf A. Heckemann, Paul Aljabar, Joseph V. Hajnal, Alexander Hammers, Jyrki Lötjönen, Daniel Rueckert. "Measurement of hippocampal atrophy using 4D graph-cut segmentation: Application to ADNI". *NeuroImage*, 52(1):109-118, 2010

Abstract

This chapter describes a new method of measuring atrophy of brain structures by simultaneously segmenting longitudinal magnetic resonance (MR) images. In this approach a 4D graph is used to represent the longitudinal data: edges are weighted based on spatial and intensity priors and connect spatially and temporally neighboring voxels represented by vertices in the graph. Solving the min-cut/max-flow problem on this graph yields the segmentation for all timepoints in a single step. By segmenting all timepoints simultaneously, a consistent and atrophy-sensitive segmentation is obtained. The application to hippocampal atrophy measurement in 568 image pairs (baseline and

month 12 follow-up) as well as 362 image triplets (baseline, month 12, month 24) confirms previous findings for atrophy in AD and healthy aging. Highly significant correlations between hippocampal atrophy and clinical variables (MMSE and CDR) were found and atrophy rates differ significantly according to subjects' ApoE genotype. Based on one year atrophy rates, a correct classification rate of 82% between AD and control subjects is achieved. Power analysis shows that 67 and 206 subjects are needed for the AD and MCI groups respectively to detect a 25% change in volume loss with 80% power and 5% significance.

5.1 Introduction

The hippocampus is one of the first structures in the brain to be affected by Alzheimer's disease [19], and hippocampal volume and especially atrophy over time has been shown to correlate with disease progression, e.g. [37, 82]. Estimates of hippocampal atrophy in longitudinal MR images can give insights into onset and progression of dementia and can serve as biomarker helping to discriminate dementia patients from healthy subjects. Since manual determination of the volume of brain structures is time-consuming and requires careful examination of intra-rater and inter-rater reliability, many efforts have been devoted to developing automated methods of atrophy rate measurement: Freeborough and Fox [57] proposed the boundary shift integral (BSI) that measures atrophy from the difference of a structure's boundaries in baseline and registered follow-up scan. SIENA is a method that quantifies atrophy from the movement of image edges between timepoints [127]. In tensor-based morphometry (TBM), the Jacobian determinants obtained from non-rigidly registering a follow-up scan to its baseline are integrated to measure atrophy [17, 93]. Alternatively, volume differences can be established by segmenting a structure of interest at different timepoints [55, 13, 109, 124]. A technique proposed by [134] that combines 3D parametric surface mapping of a structure at baseline and follow-up with automatic segmentation has recently been applied to the measurement of hippocampal atrophy in subjects from the ADNI study [109].

When measuring subtle volume changes caused by atrophy, a consistent segmentation procedure for all timepoints is crucial. Simultaneous segmentation of image sequences has been shown to increase the accuracy of atrophy measurement [156].

The majority of existing segmentation methods address single timepoints only. The method described in Chapter 3 combines graph cuts [18] and multi-atlas label propagation [73] for the segmentation of brain structures. In Chapter 4, this algorithm is embedded in a robust framework to automatically propagate a set of atlases through to a diverse image set. This chapter builds on this framework and extends the algorithm to the simultaneous segmentation of a series of MR images acquired from the same subject.

A subject-specific probabilistic atlas of a structure of interest is generated for each baseline image. After affine registration of follow-up scans to their baseline scan, this probabilistic atlas is used as spatial prior for all timepoints. This spatial prior, together with an intensity model derived from the unseen image, provides the data term to a Markov random field (MRF) which defines a graph on the image sequence connecting each voxel to a foreground and background label. To define a regularization term, additional edges between neighboring voxels within each image and between corresponding voxels along the time axis are defined. These constraints enforce a consistent segmentation within each image and across the series. Solving a single min-cut/max-flow problem on the graph defined on all timepoints yields segmentations for all images in one single step. Compared to existing methods, the additional smoothness constraint linking images along the time axis reduces the risk of spurious segmentation differences between the timepoints caused by random noise or artefacts in a particular image. Our hypothesis is that a simultaneous segmentation enables more accurate and consistent measurement of atrophy compared to segmenting the timepoints independently of each other.

The proposed method is applied to image pairs of 568 subjects from the ADNI study for whom a baseline and a month 12 follow-up scan was available. Subsequently, it was applied to the subset of 362 subjects for whom image triplets obtained at baseline,

month 12 and month 24 were available. For each series, atrophy rates were determined and its suitability as a discriminant between clinical groups was tested. Furthermore, the correlation of atrophy rates with Mini-Mental State Examination (MMSE, [54]) and Clinical Dementia Rating (CDR, [110]) scores was tested.

5.2 Materials and Methods

5.2.1 Image data

From the ADNI data described in Chapter 1.3, those subjects were used for whom a baseline and month 12 follow-up 1.5T scan were available (n=568). For 362 subjects within this population, a month 24 follow-up image was also available. All images were downloaded in April 2009. For 112 subjects, progression from MCI to AD has been reported during the study. Independently analyzed were the subject group that converted between baseline and month 12 follow-up (P-MCI$_{\leq 12}$) and the group that converted at any point after the month 12 scan (P-MCI$_{>12}$), as well as the group of subjects which had a stable diagnosis of MCI (S-MCI). While the ADNI study aims to follow all subjects for 36 months, for most subjects the examination for this timepoint was not available when this study was conducted, which means that some subjects in the S-MCI group are likely to convert to P-MCI$_{>12}$ in the future. An overview of the subject groups is given in Table 5.1: For each group the total number of subjects, number of females, and the average MMSE and CDR scores are shown, along with the development of these clinical values over time. The mean age for all subjects of 75.3±6.6 years and the mean time passed between baseline and month 12 scan of 12.96±1.32 months do not vary significantly on t-test between the groups.

Table 5.2 shows for the subset of subjects for which three timepoints (baseline, month 12, month 24) were available, the total number, number of females as well as the average change in MMSE and CDR scores between baseline and month 24. The average time between baseline and month 24 follow-up scan was 24.96±1.09 months with no significant difference between the groups. There is no significant difference for

Table 5.1: Clinical and demographical overview of the study population. Mean age of 75.3 ± 6.6 years and mean time between both scans of 12.96 ± 1.32 months for the whole population does not vary between subject groups.

	N (F)	MMSE	Δ_{MMSE}	CDR	Δ_{CDR}
CN	163 (73)	29.08±1.03	-0.07±1.39	0±0	0.02±0.19
MCI	279 (101)	27.02±1.74	-0.72±2.64	0.5±0	0.04±0.20
S-MCI	167 (60)	27.25±1.71	-0.03±2.35	0.5±0	0.02±0.15
P-MCI$_{>12}$	63 (22)	26.57±1.57	-0.97±1.95	0.5±0	0.04±0.14
P-MCI$_{\leq 12}$	49 (21)	26.88±1.89	-2.79±2.83	0.5±0	0.20±0.25
AD	126 (63)	23.48±1.85	-2.59±4.09	0.74±0.25	0.22±0.49

the clinical values at baseline and month 12 between this subset and the whole set as described in Table 5.1.

For 11 subjects in the MCI group and two subjects in the AD group, a reversion to the control and MCI group respectively has been reported. For eight and two subjects respectively, a 24 month scan is available. These subjects were excluded from the analysis.

Table 5.2: Subpopulation for which three timepoints were available. The number of subjects, number of females and average change in MMSE and CDR during 24 months are given for the six subject groups.

	N(F)	Δ_{MMSE}	Δ_{CDR}
CN	114 (54)	-0.16±1.29	0.06±0.16
MCI	165 (55)	-2.11±3.79	0.10±0.32
S-MCI	90 (29)	-0.47±2.58	0.03±0.17
P-MCI$_{>12}$	47 (16)	-4.02±4.04	0.16±0.29
P-MCI$_{\leq 12}$	28 (10)	-4.18±4.22	0.41±0.45
AD	83 (39)	-4.43±5.64	0.47±0.58

5.2.2 Hippocampus atlases

The atlases used to automatically segment the hippocampus in baseline and follow-up images are based on the hippocampal label maps provided by ADNI (Appendix A.1.1). Although other hippocampus definitions exist (e.g. [112, 67]) and can be used with the described method, the protocol used by ADNI was applied to allow better comparison with other methods.

5.2.3 4D image segmentation with graph-cuts

Building on the graph-cuts based segmentation of individual 3D MR images described in Chapter 3, this section describes an extension to 4D, dealing with sequences of MR images.

When independently segmenting every 3D image in a longitudinal sequence, the segmentation of a structure may vary between scans even if there are only small variations in the intensity [156]. This is more likely near indistinct boundaries, e.g. between hippocampus and amygdala. To be more robust against intensity variations between timepoints and against differences caused by image noise, the segmentation framework is extended from a single image to the simultaneous segmentation of a sequence of images. This is achieved by extending the graph defined by the MRF-based energy function in Equation 3.1 $E(f) = \sum_{p \in I} D_p(f_p) + \lambda \sum_{\{p,q\} \in N} V_{p,q}(f_p, f_q)$ from 3D to 4D. A 4D image I is not only defined by spatial coordinates x, y, z but also by a time coordinate t. 4D images are generated for each subject by affine registration of the follow-up scans to their baseline image, establishing correspondence between voxels in 4D. For a 4D image, a voxel $p^{x,y,z}$ has a 8-neighborhood N which incorporates the two temporally adjacent voxels $p^{x,y,z,t-1}$ and $p^{x,y,z,t+1}$ into the standard 6-neighborhood in 3D. The smoothness constraint thus applies both in space and time, and the segmentations at different timepoints are forced to be consistent in areas where only a small gray value difference between the images exists. The difference in the segmentation result of neighboring timepoints can then be expected to reflect intensity differences caused by atrophy and is less likely to be caused by noise in individual images.

Energy terms

The data term $D_p(f_p)$, consisting of a spatial prior and a probabilistic intensity model, as well as the smoothness term $V_{p,q}(f_p, f_q)$ are estimated in a similar fashion to the ones proposed for 3D segmentation in Chapter 3. To deal with a diverse set of images, the LEAP framework described in Chapter 4 is used to register multiple atlases to every baseline images in order to estimate the spatial prior. After applying LEAP, N

atlases have been registered to every image in the dataset. The spatial probability of observing a structure of interest (foreground) is determined for each voxel $p^{x,y,z,t}$ from these atlases:

$$P_A(p, f^F) = \frac{1}{N} \sum_{j=1}^{N} \begin{cases} 1, & f^F = f^{F,j} \\ 0, & \text{else} \end{cases} \tag{5.1}$$

with f^F defining the foreground label.

After affine registration of follow-up images to their baseline, the probabilistic atlas produced for the baseline image is used for all timepoints. To establish one-to-one correspondences, voxel grids of follow-up images are aligned with that of the baseline using an interpolation based on B-splines [139]. Tissue loss resulting from Alzheimer's disease can be observed as a shift of the boundaries of anatomical structures [57]. This means that differences in the segmentations of different timepoints can be expected to lie primarily in the boundary region of structures. Since the prior probability values of the atlas are low in the boundary regions, the segmentation in these areas depends mainly on the intensity model. A consistent gray value difference between two timepoints at a particular location therefore results in a segmentation difference which will be interpreted as atrophy.

To account for global intensity differences in individual scans, intensities in the follow-up scans are matched to those in the baseline scan using linear regression. A Gaussian probability distribution as the intensity model $P_F(p, f^F)$ is then estimated from all timepoints. It is defined from the voxels in the image sequence where the prior probability P_A of observing the structure of interest is at least 95%.

Simplifying from the general background model described in Chapter 3, Equation 3.3, the probability $P_B(p, f^B)$ of observing the background label f^B at a certain voxel p with intensity y_p is estimated from a mixture-of-Gaussians (MOG) model for three tissue classes (white matter (WM), gray matter (GM) and cerebrospinal fluid (CSF)). It is defined by Gaussian distribution parameters τ_k, $k = 1, 2, 3$ and previously generated and non-rigidly aligned probabilistic atlases γ_k:

$$P_B(p, f^B) = \sum_{k=1}^{3} \gamma_k P(y_p|\tau_k). \tag{5.2}$$

The smoothness term $V_{p,q}$ determining the weight of an edge connecting two voxels p, q is based on intensity differences between neighboring voxels as well as image gradients, as originally proposed in [128] and described in more detail in Chapter 3, Equation 3.5. To discriminate between spatial edges (within timepoints) and temporal edges (between timepoints), additional weighting parameter α_{spat} and α_{temp} are introduced into the MRF energy function described in Equation 3.1, allowing to give different weights to temporal and spatial edges. The 4D graph-cut model is illustrated in Figure 5.1.

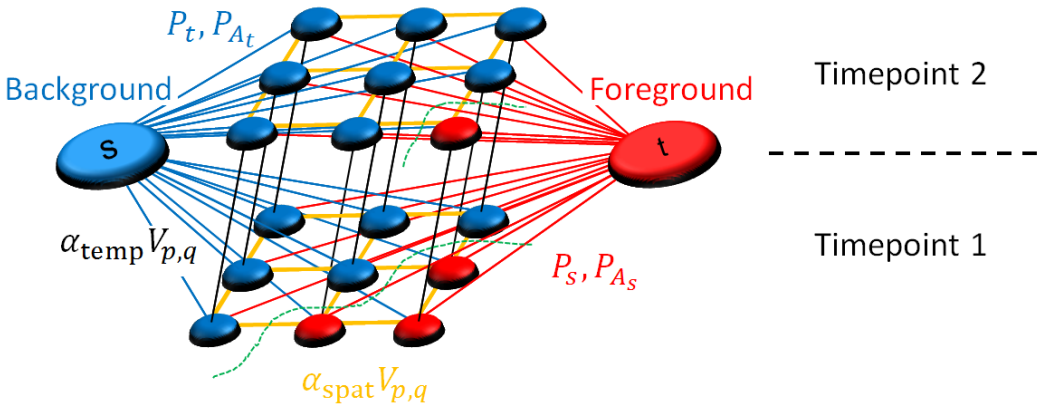


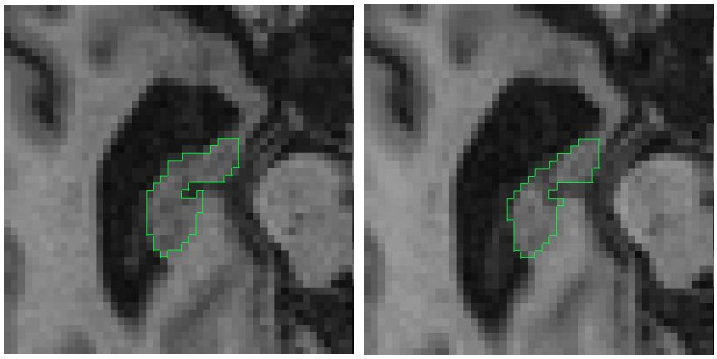
Figure 5.1: 4D graph cut segmentation: images acquired at two timepoints are connected by additional smoothness constraints (black edges) when compared to a 3D graph cut model.

5.3 Experiments and Results

The proposed 4D graph cuts method was applied to the two image sets described in Section 5.2.1: Set 1, consisting of 555 image pairs at baseline and month 12 follow-up and Set 2, consisting of 352 image triplets at baseline, month 12, and month 24.

Figure 5.2 shows a typical segmentation result for baseline and month 12 images

on a transverse section of the right hippocampus in a subject with AD. The atrophy-related discrepancy of the strong GM-CSF boundary is accurately captured and, more importantly, a consistent segmentation across timepoints is produced in areas where the hippocampus is not defined by clear boundaries.



(a) (b)

Figure 5.2: Segmentation of the right hippocampus in an AD subject. baseline (a) and month 12 follow-up (b) segmentation using 4D graph cuts.

5.3.1 Hippocampal atrophy after 12 and 24 months

Hippocampal atrophy rates in image Set 1 are shown in Table 5.3. All subject groups show a statistically significant volume loss with $p<0.001$ on a paired t-test. Mean atrophy rates (%) are shown along with the standard deviations displayed for the three clinical groups (AD, MCI, controls (CN)) as well as the different groupings of MCI subjects introduced in Section 5.2.1.

Table 5.3: Hippocampal atrophy rates (%) in 555 subjects over 12 months. Number of subjects are given in parentheses. Mean±std

	CN (163)	MCI (268)	AD (124)
r	0.78±1.77	2.19±2.88	3.82±2.25
l	0.92±1.89	2.36±2.47	3.96±2.51
r+l	0.85±1.59	2.34±2.12	3.85±1.99

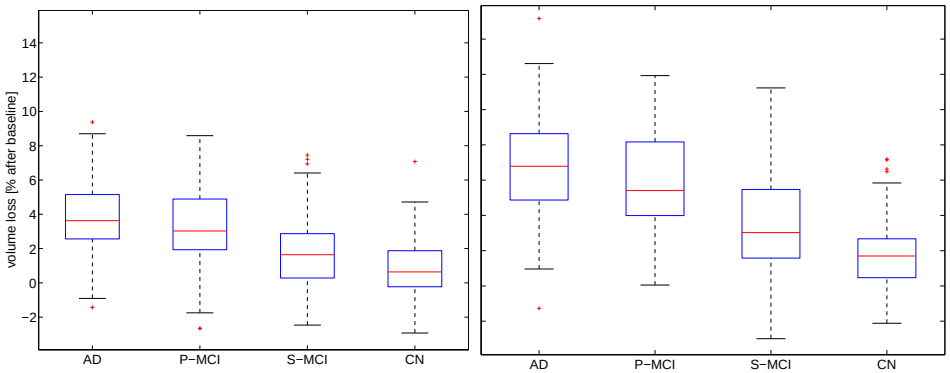
	S-MCI (156)	P-MCI (112)	P-MCI$_{\leq 12}$ (49)	P-MCI$_{>12}$ (63)
r	1.68±3.12	2.97±2.28	3.27±2.09	2.75±2.41
l	1.67±2.23	3.41±2.44	4.00±2.20	2.98±2.53
r+l	1.72±1.91	3.23±2.10	3.61±1.91	2.88±2.23

Table 5.4 shows the average atrophy rates (%) over 24 months when segmenting baseline, 12 month and 24 month images simultaneously.

Table 5.4: Hippocampal atrophy rates (%) in 352 subjects over 24 months. Number of subjects are given in parentheses. Mean±std

	CN (114)	MCI (157)	AD (81)
r	1.52±2.29	4.36±3.26	6.71±3.27
l	1.80±2.19	4.65±3.49	6.87±3.19
r+l	1.66±2.07	4.50±3.12	6.74±2.89

	S-MCI (82)	P-MCI (75)	P-MCI$_{\leq 12}$ (28)	P-MCI$_{>12}$ (47)
r	3.55±3.02	5.33±3.29	5.32±3.45	5.33±3.23
l	3.46±3.30	6.08±3.18	6.43±3.72	5.88±2.83
r+l	3.50±2.90	5.70±2.96	5.86±3.36	5.61±2.74



(a) month 12 (b) month 24

Figure 5.3: Hippocampal volume loss in % from baseline after 12 and 24 months. Box-and whisker plots for AD, P-MCI, S-MCI, CN.

Figure 5.3 shows box-and-whisker plots of atrophy rates over 12 and 24 months for normal controls, MCI converters (P-MCI), subjects with stable MCI (S-MCI), and AD subjects. The difference in atrophy rate between all clinical groups (CN, MCI, AD) is statistically significant ($p<0.001$) on a two-sample (unpaired) t-test. No significant difference was observed between P-MCI and AD, which can be explained by the fact that subjects in the P-MCI group later convert to the AD group and are therefore likely to be pathomorphologically similar.

To investigate the consistency of the proposed method as well as the influence of additional constraints when segmenting more than two timepoints, the measured atrophy obtained for the first year when segmenting two and three timepoints simultaneously was compared. T-tests indicate no significant difference between the means of matched samples ($p=0.57$). A Bland-Altman plot comparing both measures is displayed in Figure 5.4. The plot shows good agreement between the two measurements

with few outliers.

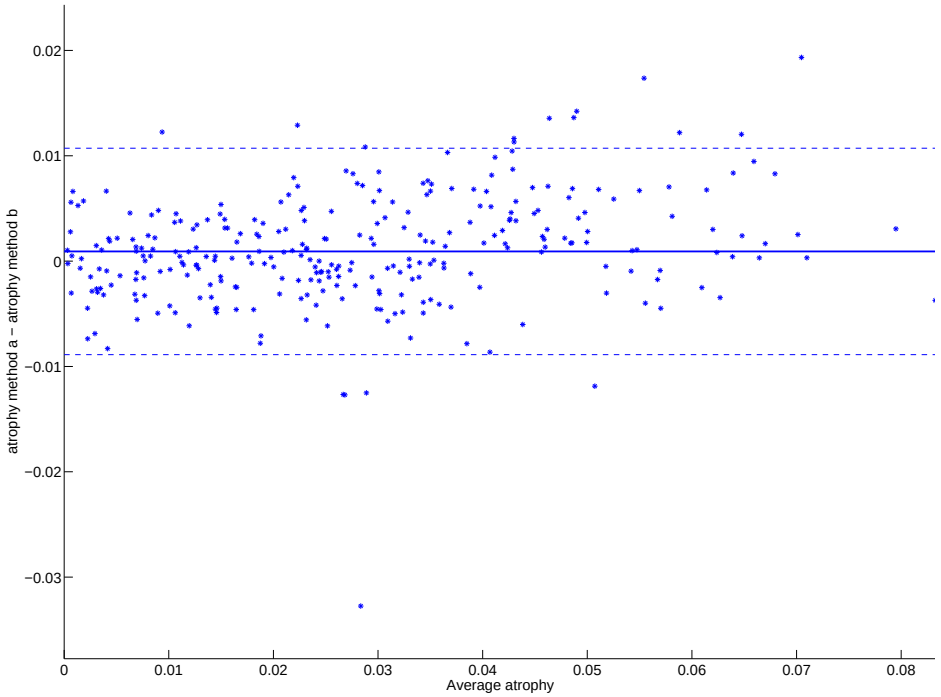


Figure 5.4: Comparison of volume loss after 12 months when segmenting two (method a) or three (method b) timepoints simultaneously. Dashed lines represent the 95% confidence interval of the mean (solid line).

5.3.2 Correlation with clinical values

Image set 1 (555 subjects with month 12 follow-up) was used to determine the correlation of atrophy with clinical data. Table 5.5 shows Pearson's r-values for the correlation of atrophy with MMSE, CDR, and the change of both values over one year. Correlations are displayed for the whole image set as well as for the clinical groups individually.

Since CDR does not vary within the MCI and CN groups at baseline, no meaningful correlation can be measured. When using all subjects, a significant correlation in the anticipated direction could be observed in all tests. Correlations were almost as strong for the MCI group and were still significant for the left hippocampus when looking at the AD group separately.

Table 5.5: Correlation of 12-month atrophy rates with clinical values. Number of subjects are given in parentheses. (a: p<0.001, b: p<0.01)

		all (555)	CN (163)	MCI (268)	AD (124)
MMSE	r	-0.43^a	-0.13	-0.31^a	-0.17
	l	-0.52^a	-0.09	-0.38^a	-0.26^b
Δ_{MMSE}	r	0.30^a	0.16	0.26^a	0.13
	l	0.36^a	0.14	0.32^a	0.22^b
CDR	r	0.38^a	N.A.	N.A.	0.14
	l	0.47^a	N.A.	N.A.	0.22^b
Δ_{CDR}	r	-0.21^a	-0.06	-0.15^b	-0.15
	l	-0.27^a	-0.08	-0.20^a	-0.23^b

5.3.3 ApoE genotype

Further tests were carried out to gain an understanding of the influence of a subjects' ApoE genotype (determined by the ApoE alleles carried) on hippocampal atrophy. Humans carry two out of three possible ApoE alleles (ε2, ε3, ε4). Carriers of ApoE4 have been shown to have a higher risk of developing AD, while ApoE2 carriers have a lower risk [92]. Table 5.6 shows the results of a two-tailed t-test comparing the atrophy rates for ε3/3 and ε3/4 carriers (ε2/2, ε2/4, ε4/4 carriers were excluded). Significant differences between the genotypes can be observed when looking at all subjects simultaneously, but also within subgroups – controls, MCI and the combination of both. No significant difference of atrophy rates in the left hippocampus can be observed when only looking at the control group.

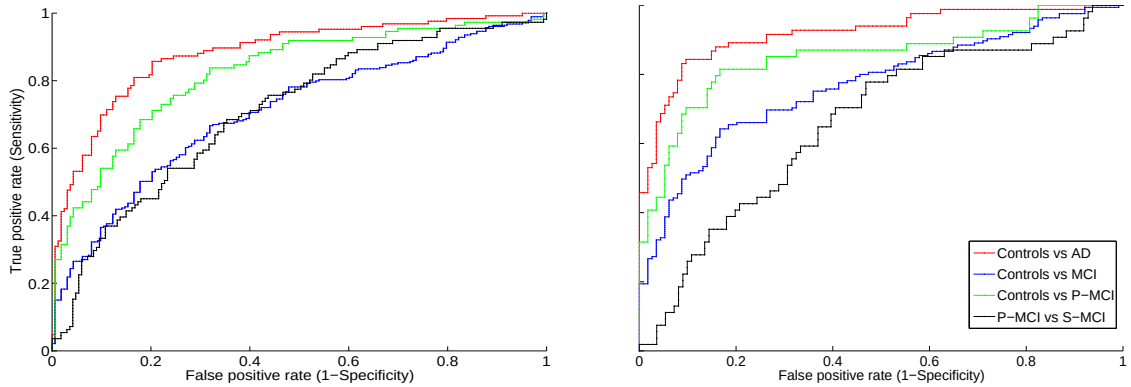
Table 5.6: T-statistics for the hypothesis of atrophy rates over 12 months in ε3/3 and ε3/4 carriers come from the same distribution. The number of subjects carrying E3 and E4 respectively is given in parentheses. a: p<0.001

	all	CN (96/42)	MCI (115/141)	CN & MCI (211/183)
r	-6.09^a	-3.21^a	-2.95^a	-5.03^a
l	-5.33^a	-1.1	-2.60^a	-4.01^a

Additionally, atrophy rates for the ε2/3 and ε3/3 carriers were compared. The only fairly significant difference in atrophy rates, however, could be observed for the left hippocampus when using all subjects (t=2.28, p = 0.02) or when combining CN and MCI groups (t=2.13, p=0.03).

5.3.4 Discrimination between clinical groups based on atrophy

Automatically determined atrophy values were tested for their power to discriminate between subject groups. Receiver operating characteristic (ROC) curves for atrophy-based classification after 12 and 24 months are displayed in Figure 5.5. The area under the curve (AUC) to classify CN vs AD, CN vs MCI, CN vs P-MCI and P-MCI vs S-MCI are 0.88, 0.71, 0.83, and 0.72 respectively when using atrophy rates measured over 12 months. Measuring atrophy over 24 months, results in values of 0.92, 0.77, 0.86, and 0.71, respectively.



month 12 month 24

Figure 5.5: ROC curves show the discrimination between subject groups. The area under the curve (AUC) for Controls vs AD, Controls vs MCI, Controls vs P-MCI and P-MCI vs S-MCI are 0.88 (0.92), 0.71 (0.77), 0.83 (0.86), and 0.72 (0.71), respectively. AUC's for rates after 24 months are given in parentheses.

A bootstrapping approach that has previously been used for classification based on hippocampal volume [29] was used to evaluate the classification rate between pairs of clinical groups: for each group 75% of the subjects were randomly selected for training. The remaining 25% were then classified according to their difference from the mean rates estimated in the training sets. The average classification rate, sensitivity and specificity for different groups after 5000 runs is displayed in Table 5.7. Values based on atrophy rates after 24 months are given in parentheses.

Using atrophy rates from the first year of observation, a classification rate of 75%-82% is obtained when discriminating between healthy controls and AD patients or

Table 5.7: Classification results using automatically determined atrophy rates after 12 months and after 24 months in parentheses.

	AD/CN	MCI/CN	P-MCI/CN	P-MCI$_{\leq 12}$/CN	P-MCI$_{>12}$/CN
Class. rate	82%(86%)	63%(72%)	76%(83%)	80%(82%)	75%(84%)
Sensitivity	81%(85%)	59%(65%)	73%(79%)	76%(69%)	72%(83%)
Specificity	83%(87%)	71%(83%)	78%(85%)	81%(86%)	75%(85%)

	P-MCI/S-MCI	P-MCI$_{\leq 12}$/S-MCI	P-MCI$_{>12}$/S-MCI
Class. rate	66%(67%)	70%(67%)	64%(68%)
Sensitivity	62%(66%)	66%(61)%	63%(70)%
Specificity	68%(69%)	72%(70)%	64%(68)%

subjects that develop AD during the study. Of clinical interest is the identification of subjects converting from MCI to AD. Early and reliable detection of these subjects could support clinical decisions for or against therapy with disease-modifying drugs. Hippocampal atrophy over the first year correctly identified 70% of subjects who converted from MCI to AD in the same period. An even more interesting result is the classification rate of 64% between subjects who did not convert within the entire observation period and subjects who converted *after* 12 months. Taking atrophy rates after 2 years, better results are achieved in all pairings except P-MCI$_{\leq 12}$ vs S-MCI.

5.3.5 Sample size calculation

For each patient group, the sample size needed in a hypothetical two-arm study to detect a reduction in the mean annual rate of atrophy was estimated. With a chosen effect size of Δ and a standard deviation σ, the following formula can be used to estimate the sample size needed:

$$n = \frac{2\sigma^2 \left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\Delta^2} \tag{5.3}$$

Following ADNI guidelines, Δ was set to 0.25μ where μ is the mean atrophy rate of the corresponding group (see Tables 5.3, 5.4). The significance level (α) was set to 0.05 and the power ($1-\beta$) to 0.8. The cutoff points of the standard normal probability distributions matching the defined significance and statistical power are $z_{1-\alpha/2} \approx 1.96$ and $z_{1-\beta} \approx 0.84$ respectively.

The total estimated sample sizes for both arms needed to detect a 25% reduction in the AD and MCI groups in intervals of 12 and 24 months are displayed in Table 5.8.

Table 5.8: Estimated sample sizes for both arms that would be needed to detect a 25% reduction in atrophy in the AD and MCI groups in intervals of 12 and 24 months.

Interval	AD	MCI
12 months	67	206
24 months	46	121

5.3.6 Segmentation accuracy

Test re-test reliability

To test the reliability of the proposed method, ten image pairs that were each acquired from the same ADNI subject in the same study session were independently segmented. When randomly selecting a reference segmentation for each pairing, the average volume difference to the second segmentation is not statistically significant[1]. The average *absolute* difference between the measurements is $1.2 \pm 1.3\%$ of their average value. Applying the presented 4D graph cuts method to these image pairs reduces the average absolute difference to $0.34 \pm 0.36\%$. This shows that segmentations obtained simultaneously from multiple time points are more consistent than single-time point segmentations.

Comparison of simultaneous to semi-automatic independent segmentation

To assess the importance of segmenting images from all timepoints simultaneously, atrophy estimates were compared to those based on the label maps provided by ADNI as described in Section 5.2.2. These label maps have previously been used to study hippocampal atrophy in work by Schuff et al. [124]. The similarity index (SI) [42] was used to measure average overlaps between the manually corrected label maps and the segmentation produced by the proposed method. The average overlap for 262 baseline

[1]The hypothesis that the distribution has zero mean can not be rejected with $p = 0.32$

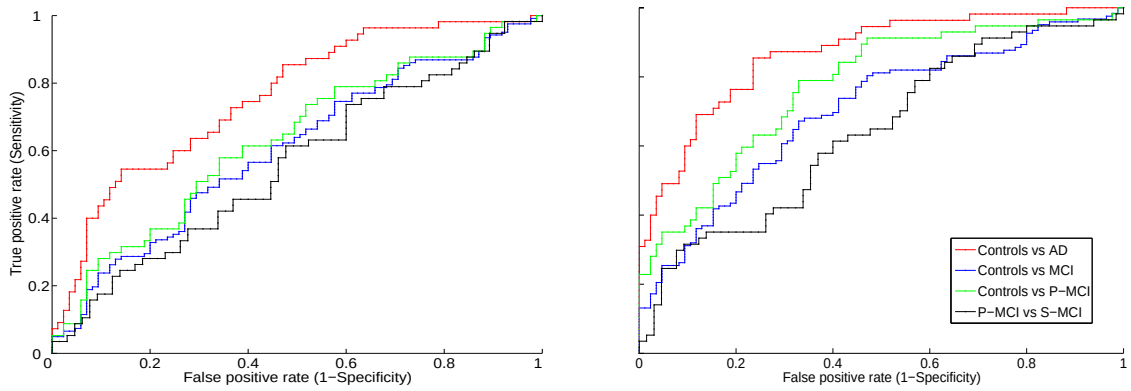
and month 12 follow-up images is 0.83±0.04. There is no significant difference between left and right hippocampus.

The subset of images for which label maps at baseline and month 12 were provided by ADNI was used to compare both approaches of atrophy measurement. Resulting atrophy rates (%) are shown in Table 5.9. Despite the significant differences in mean values, there is good correlation between both measures with $r = 0.45$ and $p < 0.001$ when looking at all values. The correlation is still high and significant ($p < 0.001$) when looking at AD and MCI groups separately ($r = 0.61$, $r = 0.42$ respectively).

Table 5.9: Average atrophy rates (%) for the subset of image Set 1 for which hippocampal label maps were provided by ADNI. Atrophy rates based on these label maps are compared to automatically determined rates based on the proposed method. Numbers of subjects are given in parentheses. mean±std

	CN (85)	MCI (122)	S-MCI (65)	P-MCI (57)	AD (55)
ADNI labels	1.10±5.82	3.23±5.58	2.72±5.49	3.81±5.66	6.27±4.84
4D graph-cuts	0.9±1.61	2.31±2.08	1.82±1.89	2.87±2.16	3.67±1.82

Figure 5.6 shows ROC curves, demonstrating the ability of both measurements to discriminate between clinical groups. Although the mean difference between clinical groups is higher with the independent ADNI label maps, classification results are better with the proposed graph-cuts approach performing simultaneous segmentation. The AUC for the classification between CN vs AD improves from 0.76 to 0.87 while the clinically most interesting classification between P-MCI and S-MCI is improved from 0.58 to 0.66. The improvement with the described method can be explained by the larger precision of the proposed method, evidenced by the lower standard deviation of the atrophy rates measured. The sample sizes required to detect a 25% reduction in atrophy in the AD and MCI groups confirm this observation with substantially lower values for the proposed method. Atrophy rates based on the label maps provided by ADNI result in samples sizes for both arms of 150 and 750 subjects for the AD and MCI groups respectively. Applying 4D graph-cuts to this subset results in reduced sample sizes of 62 and 204 subjects respectively.



ADNI label maps | 4D graph-cuts

Figure 5.6: ROC curves show the discrimination between subject groups. The area under the curve AUC for (ADNI labels/4D graph-cuts) Controls vs AD, Controls vs MCI, Controls vs P-MCI and P-MCI vs S-MCI are 0.76/0.87, 0.60/0.70, 0.63/0.77, and 0.58/0.66, respectively.

Temporal smoothness term

To assess the influence of the weighting factors α_{temp} and α_{spat} introduced in Section 5.2.3 that weights spatial and temporal edges individually, atrophy measurement over 12 months for Set 1 was performed with different parameter settings. While small parameter changes do not influence the segmentation outcome substantially, weighting spatial edges with around 20 times higher than temporal edges leads to a robust framework that results in consistent segmentations but is still flexible enough to accurately detect atrophy. Depending on the structure to be segmented and expected difference over time, temporal constraints can be varied in different settings.

Setting $\alpha_{\text{temp}} = 0$ leads to average atrophy rates of 3.94 ± 2.13, 2.32 ± 2.31, 0.87 ± 1.66 for the AD, MCI and CN groups respectively. The increased difference in mean values does not outweigh the increase in standard deviation and therefore results in slightly worse classification results and larger sample sizes needed to detect change[2].

[2]Classification was performed as described in Section 5.3.4, results are not shown here. Using Equation 5.3 shows slightly higher sample sizes compared to the ones reported in Section 5.3.5.

5.4 Discussion and Conclusion

This chapter presented a 4D graph-cut segmentation method and applied it to measuring hippocampal atrophy in longitudinal MR images from AD patients, subjects with MCI as well as age matched healthy controls. In the evaluation, 568 image pairs (baseline and month 12 follow-up) as well as 362 image triplets (baseline, month 12, month 24 follow-up) were segmented simultaneously. The resulting atrophy rates confirm previous results for hippocampal loss in AD and healthy aging, with atrophy rates significantly higher in AD (3.85±1.99 vs. 0.85±1.59). The values are in the same range as atrophy rates for both groups reported in a recently published meta-analysis of hippocampal loss rates in AD which combines nine studies using manual and automatic approaches [12]. Two recent studies report substantially different atrophy rates for a similar subset of ADNI subjects: Morra et al. [109] with AD: 5.59±7.24, CN: 0.66±5.96 and Schuff et al. [124] with AD: 4.4±5.88, CN: 0.8±5.63 [3]. While the hippocampus atlases used in the presented work are based on the same protocol used in [124], the differences to the atrophy rates reported in [109] may partly be explained by a potential difference in region definition. In addition, both previous studies report relatively large confidence intervals which make an estimate of mean values less reliable.

Atrophy rates in subjects with progressive MCI were found to be significantly higher than in subjects with a stable diagnosis of MCI. Furthermore, subjects with stable MCI show higher atrophy rates than control subjects. These results confirm findings by [147] and are also supported by the finding of significantly reduced cortical thickness in the P-MCI group compared to the S-MCI group reported in, e.g., [87]. Our results furthermore show that subjects converting to AD during the first year of the study showed significantly higher atrophy in that time period. More interesting, however, is the significantly higher atrophy rate of subjects converting to AD after year one. This suggests that substantial loss in hippocampal volume can be observed

[3] Standard deviations were calculated from 95% confidence intervals and standard errors respectively as well as sample sizes provided in the original work.

before a conversion to AD is diagnosed with psychological tests.

Automatically determined atrophy rates over 12 months were used to determine their correlation with clinical variables, comparing the achieved results to previously reported values using a similar subset of ADNI images [109]. Due to differences in the ADNI subjects used, a direct numerical comparison of both methods is not possible. Stronger and statistically more significant correlations indicate, however, that the method proposed in this work achieves better accuracy. When using all subjects, a strong and highly significant correlation between atrophy rates and MMSE, CDR as well as the change of both variables over time could be observed. Taking into account the definition of these clinical variables and the difference in atrophy reported, these correlations are as expected. When looking at the MCI group separately, the correlation is almost as significant. In the AD group, however, only a relatively poor correlation between atrophy and clinical variables was measured for the left hippocampus. This confirms findings by [109]. Apart from the lower power to detect correlation caused by the relatively small group size, a potential explanation is the heterogeneity of the AD group with respect to change in clinical variables (see Table 5.1). The absence of a significant correlation for the control group can probably be explained by the small amount of variation of both atrophy rates and clinical variables.

In further tests, the influence of a subject's ApoE gene status on hippocampal atrophy was investigated. The presented results show a statistically highly significant difference between $\varepsilon 3/3$ and $\varepsilon 3/4$ carriers when combining all subject groups. Remarkably, the difference is still significant when looking at control or MCI groups only. Only a weak significance could be measured for the difference in atrophy rates of $\varepsilon 2/3$ and $\varepsilon 3/3$ carriers.

The reported atrophy rates can be used to classify between clinical subject groups. Although to our knowledge no classification results based on hippocampal atrophy have been published for the ADNI group so far, other classifiers have been proposed. Based on baseline volume of the hippocampus, [29] report a rate of 64% for the clinically important classification between MCI-converters (P-MCI) and subjects with stable

MCI (S-MCI). In [62], a more sophisticated classifier based on hippocampal shape features achieves discrimination between MCI and controls with an accuracy of 80%.

Hippocampal atrophy rates over 12 months based on 4D graph-cuts distinguish between controls and AD or MCI with a classification rate of 82% and 67% respectively. A discrimination of MCI converters from healthy subjects and especially from MCI non-converters is of clinical importance. With the proposed method, all converters could be discriminated from controls with a rate of at least 75%. Atrophy rates over 12 months allow the identification of 70% of the subjects that convert from MCI to AD in the *same* period. The classification rate of 64% between non-converters and subjects that converted *after* month 12 shows that an indication of future conversion can be obtained before clinical tests identify the subjects as AD patients. Taking atrophy rates over two years, better results are achieved in all pairings except P-MCI$_{\leq 12}$ vs S-MCI. This can probably be partly explained by missing information about subjects that progress from the S-MCI group to the P-MCI group after 24 months. Although all subjects are followed for 36 months in the ADNI study, the final examination is not available for the majority of subjects. Some subjects are likely to convert to AD after month 24 but are assigned to the S-MCI group, which spuriously reduces the classification rates. Another factor is probably the relatively small sample size for the interval between month 12 and 24 (especially for P-MCI$_{\leq 12}$), which results in relatively large confidence intervals around the mean atrophy rate (see Table 5.4).

A high level of agreement between the individual hippocampal segmentations generated by the proposed method and semiautomatically generated reference segmentations provided by ADNI (SI 0.83 ± 0.04) was found. Atrophy rates calculated on the basis of both methods were strongly correlated. Significant differences between the two approaches are seen when the comparison is based on classification rates and statistical power: On these criteria, the 4D graph-cuts based method is clearly superior. One explanation for this superiority could be the presence of increased temporal constraints when segmenting images of all timepoints simultaneously: this leads to higher consistency within the ensembles of measurements on which the atrophy calculation is

based. A comparison of the presented 4D graph-cuts approach to different state-of-the art methods for an automated measurement of atrophy is presented in Table 9.3 in Chapter 9 of this thesis.

Chapter 6

Biomarker extraction from manifold learning

This chapter is based on:

Robin Wolz, Paul Aljabar, Joseph V. Hajnal, Jyrki Lötjönen, Daniel Rueckert. "Nonlinear Dimensionality Reduction Combining MR Imaging with Non-Imaging Information Medical Image Analysis". *Submitted*, 2011

Abstract

Going on from the traditional biomarkers presented before, this chapter describes a method based on machine learning for biomarker extraction. In a low-dimensional manifold representation of inter-subject brain variation, the manifold coordinates of each image capture information about structural shape and appearance and, when a phenotype exists, about the subject's clinical state. A novel feature of the presented framework is the incorporation of subject meta-information into the manifold learning step. Information such as gender, age or genotype is often available in clinical studies and can inform the classification of a query subject. Such information, whether discrete or continuous, is used as an additional input to manifold learning, extending the Laplacian eigenmap objective function and enriching a similarity measure derived from pairwise image similarities. The biomarkers identified with the proposed method are

data-driven in contrast to a-priori defined biomarkers derived from, e.g., manual or automated segmentations. They form a unified representation of both the imaging and non-imaging measurements, providing a natural use for data analysis and visualization. The described method is tested using ApoE genotype, the CSF-concentration of $A\beta_{42}$ as non-imaging metadata and hippocampal volume as a derived imaging-biomarker for subject classification. Achieved classification results compare favorably to what has been reported in a recent meta-analysis using established neuroimaging methods.

6.1 Introduction

Like the measurements extracted with the methods presented in Chapters 3, 4 and 5, many of the well-established biomarkers for dementia from magnetic resonance (MR) images are based on traditional morphometric measures, such as the volume or shape of brain structures [55, 49, 124, 30] and their change over time [57, 127, 17, 93]. More recently, models based on machine learning techniques have been proposed which seek discriminating features over the whole brain or within a defined region of interest [51, 50, 144, 62]. Finding a low-dimensional representation of complex and high-dimensional data is a central problem in machine learning and pattern recognition. Many methods have been proposed to learn the underlying low-dimensional space of intrinsically low-dimensional data lying in a high-dimensional space. Linear models like principal component analysis (PCA) [84] and multi-dimensional scaling (MDS) [36] have been extensively used for dimensionality reduction. More recently non-linear methods like ISOMAP [133], locally linear embedding (LLE) [117] and Laplacian eigenmaps (LE) [14] have been proposed to better model highly non-nonlinear data. A more detailed overview on manifold learning techniques is given in Chapter 2.

Aljabar et al. [2] applied spectral analysis [145] to pairwise label overlaps obtained from a structural segmentation to discriminate AD patients from healthy controls. Focusing on intensity-based similarities between MR brain images, Klein et al. [88] used vectors defined by the similarities of a given query subject with a cohort of training

images as features from which to learn a classifier. Computer vision applications, particularly for face recognition, also use pairwise similarities to learn a low-dimensional subspace and to classify unseen images mapped to this space [24, 72, 157]. These methods are typically linear, making it easy to transform data from image space into the learned subspace, but this linearity can limit the ability to generalize to complex datasets. Indeed, recent work suggests that the complex natural variation of brain images is best described by nonlinear models [63, 66]. This chapter aims to learn the manifold[1] structure of brain images in healthy ageing and neurodegeneration by considering both clinically labeled and unlabeled image data.

Nonlinear dimensionality reduction of a set of brain images with Laplacian eigenmaps (LE) is based on pairwise image similarities that can be evaluated either over the whole image or in a region of interest (ROI). A weighted similarity graph is built that represents neighborhood information in the image data set. With the Laplacian of the graph, a low-dimensional embedding that respects the input relations is determined. The LE objective function, which is based on edge weights in the similarity graph, places more similar images in the input space closer in the embedded space. Building on this principle, a method is proposed that extends the LE objective function in order to learn a manifold not only defined by pairwise image similarities but also by some metadata available for the subjects under consideration. Such metadata in a clinical setting can be discrete (e.g. gender) or continuous (e.g. age). The described method extends the similarity graph defined in LE by a set of additional nodes representing a number of discrete states or intervals of a continuous variable. The weights from every subject to these nodes are defined based on the subjects' metadata. This approach groups subjects with similar metadata closer in the manifold. The proposed method is related to previous work where binary label information in partially labeled data sets is used to enforce constraints in a low-dimensional manifold representation [34] . Optimizing the extended LE objective function, results in an embedding that incorporates metadata and pairwise image similarities at the same time. The coordinates

[1]Here, the terms "manifold learning/embedding" and "dimensionality reduction" are used interchangeably.

of a particular subject in the low-dimensional space can then be regarded as encoding information about the shape and appearance of the brain as well as the state of the meta-variable and thus about clinically relevant differences across the population described by both measures. Images with clinical labels can be used to infer information about unlabeled images in their neighborhood within the learned geometrical space.

Support vector machines (SVM) are used to perform classification of unlabeled subjects in the low-dimensional manifold. Furthermore, the power of the manifold representation to predict clinical variables by fitting a multiple linear regression model of clinical data versus manifold coordinates is tested.

The contribution of this chapter can be summarized as follows: a method for the extraction of a unified biomarker combining imaging information with non-imaging metadata is described. Such a unified representation makes its use for data analysis and for visualization in a potential clinical application more powerful. The method can handle discrete and continuous metadata and offers a natural way to deal with incomplete information. Compared to classical biomarkers, the proposed method is data-driven and only requires minimal a-priori information. In the evaluation, the 420 ADNI subjects are used for which a measurement of the cerebrospinal fluid (CSF) -concentration of the Aβ_{42} protein and the subject's ApoE genotype were available. In addition to these meta variables, the power of automatically derived hippocampal volumes as derived MR-based meta-information is tested as well as the ability of the proposed method to combine different metadata in a single manifold learning step.

6.2 Method

6.2.1 Manifold learning using pairwise image similarities

A set of images $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\} \in \mathbb{R}^D$ is described by N images $\mathbf{x}_i$, each being defined as a vector of intensities, where D is the number of voxels per image or region of interest. Assuming $\mathbf{x}_1, ..., \mathbf{x}_N$ lie on or near an d-dimensional manifold $\mathcal{M}$ embedded in $\mathbb{R}^D$, a low dimensional representation $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\}$ with $\mathbf{y}_i \in \mathbb{R}^d$ of the input

images in $\mathcal{M}$ may be learned. From the available dimensionality reduction techniques (see Chapter 2.4), Laplacian eigenmaps (LE) [14] are used to be able to directly deal with image similarities in contrast to distances. The embedding function is

$$f : \mathbf{X} \to \mathbf{Y}, \quad \mathbf{y}_i = f(\mathbf{x}_i).$$

An undirected weighted graph $G = \langle V, E \rangle$ with N nodes V representing the images and edges E connecting neighboring nodes is defined on $\mathbf{X}$. Edge weights are defined based on pairwise image similarities s_{ij}. A k-nearest neighbor graph is defined, where the weight $w_{ij} = s_{ij}$ if $\mathbf{x}_i$ is among the k nearest neighbors of $\mathbf{x}_j$ or vice versa and $w_{ij} = 0$ otherwise. The w_{ij} are combined to form the weight matrix $\mathbf{W}$. Image similarities s_{ij} are typically based on intensity differences or a deformation-based metric either of which may be evaluated over the whole brain or in a region of interest. A low-dimensional representation $\mathbf{y}_i = f(\mathbf{x}_i)$ that respects the pairwise similarities w_{ij} can be obtained by minimizing the energy function

$$\sum_{ij} (\mathbf{y}_i - \mathbf{y}_j)^2 w_{ij}. \tag{6.1}$$

This energy function ensures that more similar images in the input space are closer together in the embedded space. With the diagonal degree matrix $\mathbf{D} = \sum_j w_{ij}$, this can be reformulated as

$$\begin{aligned} \sum_{ij} (\mathbf{y}_i - \mathbf{y}_j)^2 w_{ij} &= \sum_{ij} \left(\mathbf{y}_i^2 + \mathbf{y}_j^2 - 2\mathbf{y}_i\mathbf{y}_j\right) w_{ij} \\ = \sum_i \mathbf{y}_i^2 D_{ii} + \sum_j \mathbf{y}_j^2 D_{jj} - 2\sum_{ij} \mathbf{y}_i\mathbf{y}_j w_{ij} &= 2\mathbf{y}^T\mathbf{L}\mathbf{y} \end{aligned} \tag{6.2}$$

with the graph Laplacian $\mathbf{L} = \mathbf{D} - \mathbf{W}$. Since $\mathbf{L}$ is positive semidefinite, the minimization problem can be formulated as

$$\underset{\substack{\mathbf{y} \\ \mathbf{y}^T\mathbf{D}\mathbf{y}=1}}{\operatorname{argmin}} \mathbf{y}^T\mathbf{L}\mathbf{y} \tag{6.3}$$

where the constraint $\mathbf{y}^T\mathbf{D}\mathbf{y}$=1 removes an arbitrary scaling factor in the embedding and prevents the trivial solution where all y's are set to zero [14]. Finding the y_i's that optimize this objective function can be formulated in closed form as finding the eigenvectors associated with the d smallest non-zero eigenvalues of the generalized eigenvector problem

$$\mathbf{L}\nu = \lambda\mathbf{D}\nu. \tag{6.4}$$

6.2.2 Manifold learning incorporating non-imaging information

In many settings, an additional variable z_i providing further information on subject i may be available in addition to MR imaging data. Such meta-information can inform judgments such as clinical diagnosis. A method is proposed to incorporate such information into the manifold learning process (Section 6.2.1). The hypothesis is that by using this additional information, a more accurate representation of the population can be obtained leading to a more reliable biomarker in the low-dimensional space. Metadata available in a clinical setting can be defined by discrete categories (two or more), or by a continuous variable. Examples of discrete variables are gender, blood- or genotype. Continuous variables can be, e.g., a subject's age, weight or measurements derived from a phenotype associated with the disease of interest. In [34], a graph G describing the LE objective function in Equation 6.1 is extended by two nodes, each representing one of two classes available for training data. Connecting each training subject with its respective class node imposes the class differences in the training data on the manifold structure. Assuming generalizability between labeled and unlabeled nodes, a more accurate classification performance on the test data is expected. Extending this idea, the proposed method uses metadata available for *all* or a *subset* of subjects in contrast to the class labels itself to enrich the low-dimensional representation. The proposed framework supports to incorporate metadata from one or more measures, each defining M discrete classes or a continuous interval leading to a fuzzy-class membership.

Graph G is extended by M nodes $\hat{V}$ representing the metadata variable z and called *support nodes* in the following. By connecting each image $\mathbf{x}_i$ to the support nodes with weights defined according to the value of z_i, the distribution of the meta-variable will influence neighborhoods in the low-dimensional representation. In the discrete setting with $z_i \in Z_d = \{z^1, ..., z^M\}$, the weight $\hat{w}_{im}$ between subject i and the m^{th} support node is defined by

$$\hat{w}_{im} = \begin{cases} 1, & \text{if } z_i = z^m \\ 0, & \text{otherwise.} \end{cases} \tag{6.5}$$

In order to map a continuous metadata variable $z_i \in Z_c = \left[z^a, z^b\right]$ to a discrete number of support nodes, the input space is subdivided into M subintervals $\bar{z}^m \in Z_c^m = \left[z^{a,m}, z^{b,m}\right], k = 1, ..., M$. Each of these subinterval is then represented by a support node in graph G. The bounds of the M subintervals are defined as

$$\begin{aligned} z^{a,m} &= P_z\left((m-1)\frac{100}{M}\right) \\ z^{b,m} &= P_z\left(m\frac{100}{M}\right) \end{aligned} \tag{6.6}$$

where $P_z(x)$ gives the x^{th} percentile of interval z. With the mean value $\mu^m = \frac{1}{|Z^m|}\sum_{z \in Z^m} z$ of interval $\bar{z}^m$, the continuous weight $\hat{w}_{im}$ between subject i and the m^{th} support node is defined based on the distance between z_i and μ^m, grouping subject i closer to subjects with a similar value of z:

$$\hat{w}_{im} = \begin{cases} \frac{1}{c}\left(1 + (z_i - \mu^m)^2\right)^{-1} & \text{, if } z_i \text{ is available} \\ 0 & \text{, otherwise.} \end{cases} \tag{6.7}$$

where c is a normalising constant to ensure $\sum_m \hat{w}_{im} = 1$. The weighting schemes in the discrete and continuous settings for the case where an additional variable z is available for all images are illustrated in Figure 6.1. Incorporating the M support nodes $\hat{V}$ and the weights $\hat{w}_{im}$, leads to an extended Laplacian eigenmaps (E-LE) objective function

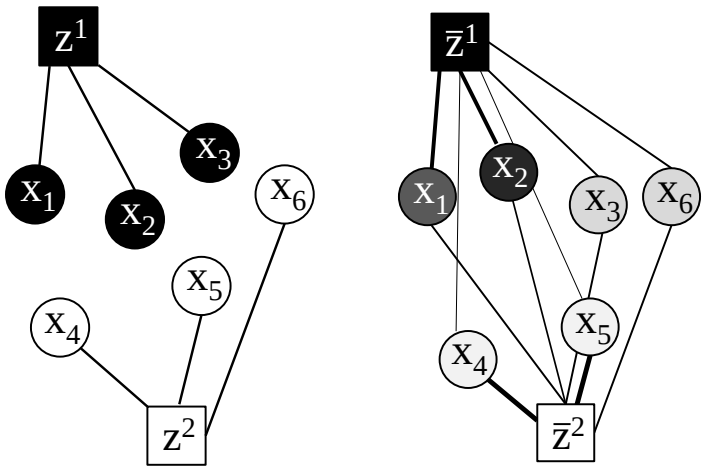


Figure 6.1: Weights defined between image nodes $\mathbf{x}_i$ and support nodes representing metadata Z. In the discrete setting (left), equally weighted edges are defined according to Equation 6.5. In the continuous setting (right), weights to both additional nodes are defined according to Equations 6.6 and 6.7. A higher weight is illustrated by a thicker edge.

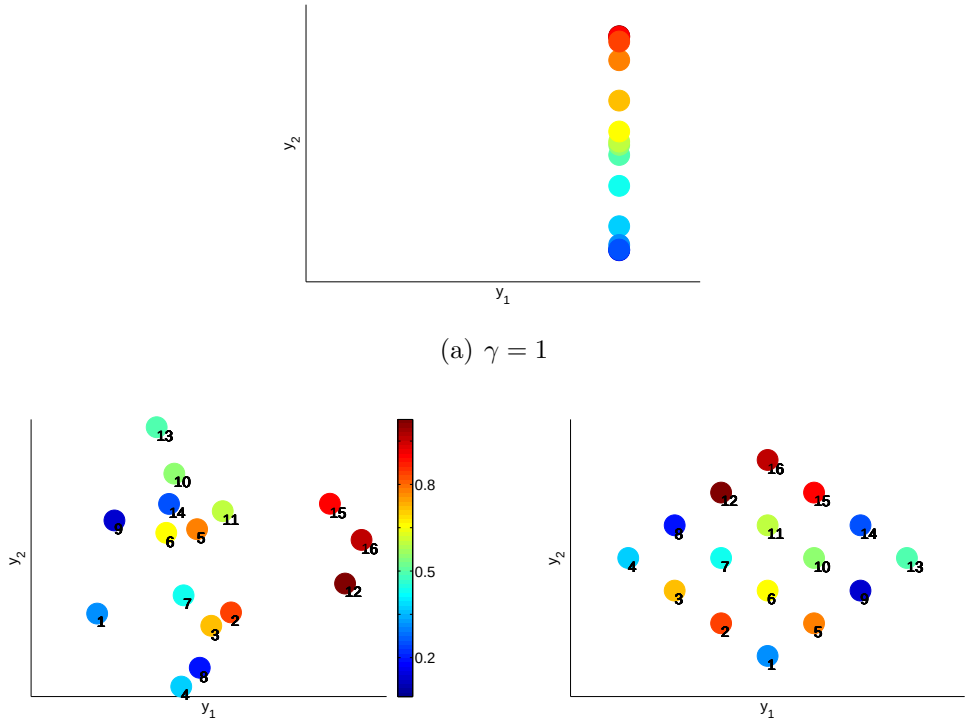
$$\gamma \sum_{ij} (\mathbf{y}_i - \mathbf{y}_j)^2 w_{ij} + \sum_{im} (\mathbf{y}_i - \hat{\mathbf{y}}_m)^2 \hat{w}_{im} \tag{6.8}$$

with $\hat{\mathbf{y}}_m$ representing the cluster center of state z^m of a discrete variable or of the interval $\bar{z}^m$ of a continuous meta-variable. The extended low-dimensional embedding space is described by

$$\mathbf{Y}' = \{\hat{\mathbf{y}}_1, ..., \hat{\mathbf{y}}_M, \mathbf{y}_1, ..., \mathbf{y}_N\} \quad , \quad \hat{\mathbf{y}}_m, \mathbf{y}_i \in \mathbb{R}^d. \tag{6.9}$$

In this embedding, the proximity of subject i to the m^{th} group (discrete or continuous) and its centroid $\hat{\mathbf{y}}_m$ is determined by the weights $\hat{w}_{im}$ defined by the metadata as well as image-based weights w_{ij} . A low weight of parameter γ arranges the subjects mainly according to the metadata, whereas a high value of γ is closer to the standard embedding with Laplacian eigenmaps which considers only pairwise image similarities. The influence, γ has on the embedding is illustrated in Figure 6.2. In a synthetic example, pairwise similarities for 16 nodes are defined from a set of distances between points in 2D to arrange them in a grid-shaped embedding when using standard Laplacian

eigenmaps. Every node is associated with a randomly assigned meta-variable varying between zero and one which is encoded by the color in Figure 6.2. In panel (a) with $\gamma = 1$, the embedding is dominated by the value of the meta-variable. Panel (b) shows an embedding influenced by both measures and panel (c) shows an embedding close to the one obtained with LE for $\gamma = 50$.



(a) $\gamma = 1$

(b) $\gamma = 8$

(c) $\gamma = 50$

Figure 6.2: First two embedding coordinates with varying influence of γ. A high weight leads to an embedding similar to the one obtained with classic Laplacian eigenmaps (c). A very low weight embeds the images mainly based on metadata (a).

With the $N \times M$ matrix $\hat{\mathbf{W}}$ defining the weights between subject i and the M support nodes, an extended weight matrix

$$\mathbf{W}' = \begin{pmatrix} \mathbf{I} & \frac{1}{2}\hat{\mathbf{W}}^T \\ \frac{1}{2}\hat{\mathbf{W}} & \gamma\mathbf{W} \end{pmatrix} \tag{6.10}$$

based on the objective function in Equation 6.8 is derived, where $\mathbf{I}$ is an $M \times M$ identity matrix. Following Equations 6.2 and 6.4 to solve the generalized eigenvector problem associated with the extended weight matrix, yields the embedding coordinates which

optimize the objective function in Equation 6.8.

6.2.3 Extraction of biomarkers

Assuming the pairwise similarities s_{ij} and the metadata variable z_i represent clinically relevant differences between relevant clinical groups, a subject's manifold coordinates $\mathbf{y}_i$ can be used as a biomarker to support inferences about their clinical state.

Classification

When aiming at classifying unlabeled subjects for which no clinical label is available, information from labeled subjects can be used to make a decision. Please note that "unlabeled" in this context refers to the clinical label (e.g. AD, healthy control) that is to be predicted. Every subject (labeled or unlabeled) may or may not have metadata associated with it that can be used to enrich the manifold learning as described in Equations 6.5 and 6.7.

When dealing with a two-class problem, the coordinates of N' labeled training images $\{\mathbf{y}_j, d_j\}, j = 1, ..., N' < N, \mathbf{y_j} \in \mathcal{R}^d$ with clinical labels $d_j \in \{-1, 1\}$ is used to train a classifier on the derived manifold coordinates $\mathbf{y}_j = y_{j1}, ..., y_{jd}$. Support Vector Machines (SVMs) minimize a Lagrangian energy function which leads to the hyperplane

$$\mathbf{a} \cdot \mathbf{y} - \mathbf{b} = \mathbf{0} \tag{6.11}$$

in the manifold space that best separates the two subject groups [22]. The location of embedding coordinates of the $N - N'$ unlabeled images in relation to this plane can then be used to classify them. While SVMs provide a natural way to separate the learned manifold space, other classifiers based on, e.g., a Linear Discriminant Analysis [90] or a k-nearest neighbor classifier [35] could be applied.

Regression

A continuous assignment can be achieved by, e.g., building a linear regression model between a clinical variable $\hat{d}_j$ versus manifold coordinates $y_{j1}, ..., y_{jd}$:

$$\hat{d}_j = a_0 + \sum_{i=1}^{d} a_i y_{ji} \tag{6.12}$$

Learning such a model from a subset of subjects for which clinical labels exist, allows its application to unlabeled subjects and predictions to be made about clinical information associated with those subjects.

6.3 Data and Results

The proposed method is evaluated by performing classification in the learned manifold space between AD patients, subjects with MCI and healthy controls. Furthermore, the power to predict clinical variables is evaluated by performing regression versus the score of a Mini-Mental State Examination (MMSE) [54]. Incorporating relevant subject-information in the form of metadata is expected to better model the difference between two populations, leading to an improved classification and regression performance. Finally, the performance of the described method in comparison to other approaches that combine different measures to perform classification is inspected.

6.3.1 Subjects

ADNI provides the ApoE genotype (determined by the ApoE alleles carried) for all subjects. Humans carry two out of three possible ApoE alleles ($\varepsilon 2$, $\varepsilon 3$, $\varepsilon 4$). Carriers of $\varepsilon 4$ have been shown to have a higher risk of developing AD, while $\varepsilon 2$ carriers have a lower risk [92]. In addition an Aβ_{42} protein analysis of cerebrospinal fluid (CSF) is available for a subset of ADNI subjects. A decrease in the concentration of this protein has been shown to be associated with a development of AD [138]. In this chapter, the 1.5T T1-weighted baseline images of the 420 subjects for which a CSF analysis was

available were used. Out of 201 MCI subjects, 89 were progressive, i.e. were diagnosed as converting to AD as of October 2010. Stable (S-MCI) and progressive (P-MCI) subjects where therefore analyzed independently[2]. Table 6.1 presents an overview of the subjects studied and their metadata as well as their MMSE scores used for clinical diagnosis.

Table 6.1: Subject data of the study subjects are shown for the different groups. Non-imaging metadata in the form of ApoE genotype and Aβ_{42} concentration as well as the derived imaging metadata, hippocampus volume, are presented. Carriers of the ApoE $\varepsilon 2/\varepsilon 4$ alleles are shown. The remaining subjects only carry the $\varepsilon 3$ allele. There is no significant difference in age between the clinical groups with an average age of 74.95±7.03 years.

	Subject Data		Non-Imaging metadata		Derived metadata
	N (F)	MMSE	$\varepsilon 2/\varepsilon 4$	Aβ_{42} (pg/ml)	Hippo. Vol. (cm^3)
CN	116 (56)	29.12 ± 1.02	16/28	202.3 ± 57.5	4.53 ± 0.55
S-MCI	112 (36)	27.16 ± 1.75	9/49	178.9 ± 61.6	4.26 ± 0.59
P-MCI	89 (33)	26.64 ± 1.8	1/52	146.3 ± 46.30	3.93 ± 0.65
AD	103 (43)	23.55 ± 1.87	4/63	147.5 ± 45.8	3.92 ± 0.73

6.3.2 Pairwise image similarities

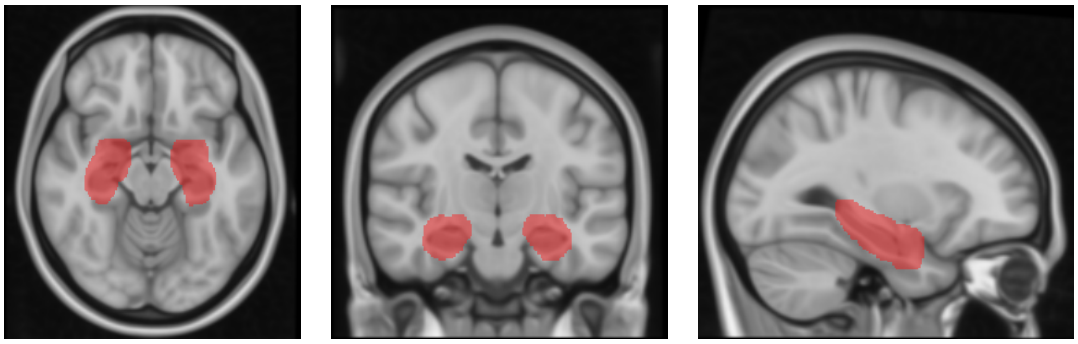
To measure pairwise image similarities, all 420 study images were aligned to the MNI152-brain T1 atlas [106] using a coarse non-rigid registration modeled by free-form deformations (FFDs) with a 10mm B-spline control point spacing [118] between the corresponding image and the atlas. A *coarse* non-rigid registration allows alignment of structures of interest while retaining inter-subject variation to measure image similarities. While the proposed framework is general and, for example, allows the use of a deformation-based metric, an intensity-based similarity measure was selected for this work. This choice was based on the expectation of only relatively subtle difference between individual images in a defined region of interest. Cross correlation (CC) between pairs of images $\mathbf{x}_i$ and $\mathbf{x}_j$, is used to specify the similarity s_{ij} defining the weight w_{ij} for manifold learning when optimizing the objective functions in Equations 6.1 and 6.8. CC was selected in favor of an entropy-based measure like mutual

[2]Note that since the ADNI study is still ongoing it is likely that some subjects will convert from the S-MCI group to the P-MCI group in the future.

information [130] since all images are based on the same modality (MRI) and a linear relationship between image intensities is expected. CC is defined as:

$$s_{ij} = \frac{\sum_k (x_{i,k} - \bar{x}_i)(x_{j,k} - \bar{x}_k)}{\sqrt{\sum_k (x_{i,k} - \bar{x}_i)^2 \sum_k (x_{j,k} - \bar{x}_j)^2}} \tag{6.13}$$

where $x_{i,k}$ denotes the gray value at the k^{th} voxel and $\bar{x}_i$ denotes the mean gray value of image i. The medial temporal lobe and in particular hippocampus and amygdala have been shown to be predominantly affected by onset and progression of MCI and AD [45]. The evaluation of pairwise similarities were therefore restricted to a region defined around both structures in the template space (see Figure 6.3).



(a) Transverse (b) Coronal (c) Sagittal

Figure 6.3: Orthogonal views of MNI152 space showing the ROI around hippocampus and amygdala used to evaluate pairwise image similarities.

6.3.3 Experiments

Based on the objective function given in Equation 6.1, traditional Laplacian eigenmaps was applied to obtain a low-dimensional representation of all 420 study images using image similarities s_{ij} only. The neighborhood size k used to define graph G did not substantially influence results when varying between 10 and 50 and was set to $k = 20$ following results presented in [63]. In addition to classic LE, the extended objective function proposed in Equation 6.8 was used to incorporate both discrete and continuous metadata into the manifold learning process. ApoE genotype and Aβ_{42}

concentration were used as clinical, non-imaging information[3]. In addition, automatically determined hippocampal volumes extracted with the LEAP framework described in Chapter 4 were used as a derived imaging biomarker to enrich the manifold learning process. Average hippocampal volumes (right + left) for the different subject groups are displayed in the very right column of Table 6.1. Furthermore, the impact of adding support nodes to more than just one meta-variable was evaluated. In particular, the combination of CSF with hippocampal volume and CSF with hippocampal volume and ApoE genotype were tested. The following list gives an overview of the different experiments performed:

I : Laplacian eigenmaps (LE)

II : Extended LE (E-LE) with ApoE genotype

III : E-LE with Aβ_{42}

IV : E-LE with hippocampal volume

V : E-LE with Aβ_{42} and hippocampal volume

VI : E-LE with Aβ_{42}, hippocampal volume and ApoE genotype

For the discrete variable in experiment II, ApoE genotype, $M = 3$ support nodes are defined, each trivially associated with a possible genotype (z_1: subjects that carry at least one $\varepsilon 2$ allele. z_2: subjects that carry at least one $\varepsilon 4$ allele. z_3: subjects that only carry the $\varepsilon 3$ allele). Following Equation 6.5, $\hat{w}_{im}$ is set to one if subject i has a genotype associated with node m, otherwise it is set to zero. For the continuous variables in experiments III and IV, Aβ_{42} concentration and hippocampal volume, a continuous weighting $\hat{w}$ is defined as described by Equations 6.6 and 6.7. To accommodate the four clinical groups (CN, S-MCI, P-MCI, AD), $M = 4$ support nodes were used with subintervals $\bar{z}^m, m = 1, ..., 4$ to describe the metadata as defined in Equation 6.6. For experiments V and VI, that use more than one meta-variable, edges to the

[3]It should be noted that neither variable is part of the inclusion / exclusion criteria for the different clinical diagnoses defined by the ADNI study [111].

support nodes associated with all variables are defined. This results in $M = 8$ and $M = 11$ support nodes for experiments V and VI respectively with weights defined as in experiments II-IV. Before performing the experiments described below, all embedding coordinates were corrected for subject age using a multiple linear regression model. Figure 6.4 shows exemplars of the projected embedding onto the first two coordinate directions when using standard Laplacian eigenmaps (top panel) and the proposed method with hippocampal volume as metadata (bottom panel). A separating hyperplane between AD and control subjects as defined by SVM is displayed in both cases. Better discrimination between the two groups can be observed when using the proposed method especially for subjects close to the separating plane.

6.3.4 Parameter settings

Dimension d

The selection of the optimal number of embedding coordinates d is not an obvious task. For different applications in manifold learning of brain images [63, 151, 83, 66, 150], different numbers of dimensions have been shown to produce good results. To get an overview of how the proposed classification framework reacts to varying the dimension of the manifold, classification between clinical groups was performed on the 418 ADNI baseline images not used in the evaluation (subjects for which no CSF information is available). Classification accuracy was evaluated for the pairings AD vs CN, P-MCI vs S-MCI and P-MCI vs CN when varying the dimension $d \in [1, \ldots, 50]$. To get a more robust measure of the optimal embedding dimension, the average accuracy was evaluated for 10 bins, each covering 5 dimensions. Average classification results and standard deviation for the 10 bins are displayed in Figure 6.5 (a). A clear improvement of classification accuracy can be observed when increasing the dimension from bin one, $d \in [1, 5]$, to bin two, $d \in [6, 10]$. From the fourth bin covering $d \in [16, 20]$, the classification accuracy decreases. Following these results, classification performance was evaluated in all experiments for $d \in [6, 15]$ and average classification rates are reported.

Weighting factor γ

The weighting factor γ defined in Equations 6.8 and 6.10 determines how much the final embedding is influenced by image similarities s_{ij} and metadata z_i. To evaluate the influence it has on classification accuracy, the performance on the images for which no CSF measurement is available (N=418) was evaluated when exemplarily using hippocampal volume as metadata. Figure 6.5 (b) shows classification results averaged over dimensions $d \in [6, 15]$ plotted over varying γ. As γ is increased, initially improved results finally asymptote to the results obtained with standard LE (illustrated with the red line in Figure 6.5 (b)). Following these results, the parameter was set globally to $\gamma = 8$ for all experiments described below and all types of metadata. Tuning γ individually for different types of metadata is expected to further improve results but requires a more complex training and makes the application to new datasets more difficult.

6.3.5 Classification

The manifold representations, obtained from standard LE embedding and from the five experiments using the extended version (E-LE) proposed in this work, are used to perform classification between the different clinical subject groups. For each relevant pairing (AD vs CN, S-MCI vs P-MCI, CN vs P-MCI), a leave-25%-out cross-validation was performed using the image set described in Section 6.3.1 that was not used for parameter setting as described in Section 6.3.4. Average classification rates after 1,000 runs are determined for every dimension $d \in [6, ..., 15]$. To arrive at a more robust and generalizable result, average classification rates over these 10 dimensions are reported.

Table 6.2 presents the correct classification rates for all six experiments. For each experiment, the multiple runs provide a distribution estimate for the corresponding classification rate[4]. For each pair of clinical groups, these distributions were used to carry out unpaired t-tests between the results of methods I (LE) and IV (E-LE with ApoE, hippo. vol. and Aβ_{42}) with the respectively remaining methods in order to esti-

[4]All estimated distributions passed a normality test using a Kolmogorov-Smirnov test at $\alpha = 0.05$.

mate the significance of any performance improvements when incorporating metadata. Resulting p-values are presented in Table 6.2. For comparison, correct classification rates when only using the different sources of meta-information are presented in the bottom part of Table 6.2.

6.3.6 Regression

The Mini-Mental State Examination is a psychological test to screen for cognitive impairment. To test the ability of manifold coordinates to predict clinical variables, a multiple linear regression model of MMSE was fitted versus manifold coordinates. A model that predicts MMSE from the first d manifold coordinates was used:

$$\text{MMSE} = a_0 + \sum_{i=1}^{d} a_i y_i. \tag{6.14}$$

The model was evaluated for $d = 15$ and for comparison with work presented in [63] also for $d = 1$. Regression results for both dimensions using the manifold representations from experiments I-VI, are displayed in Table 6.3.

6.3.7 Alternative approaches to incorporate metadata

There are several alternative approaches to perform classification based on multiple measurements or to learn a manifold based on more than one similarity measure. In this section, two obvious choices of alternatives are considered:

Concatenation of feature vectors in a SVM-based classification

When performing SVM-based classification, an extended feature vector $\mathbf{f}_i = \{\mathbf{y}_i, z_i\}$ concatenating the manifold coordinates of subject i, $\mathbf{y}_i$ with its meta-variable z_i can be defined. Classification can then be performed in the resulting $d + 1$ dimensional space. Table 6.4 shows classification results using this approach for ApoE genotype, Aβ_{42} concentration and hippocampal volume. In addition, p-values for the differences between these results and the relevant results in Table 6.2 are presented.

Table 6.2: Correct classification accuracy (ACC), sensitivity (SEN) and specificity (SPE) (%) for classic Laplacian eigenmaps (LE, I) and the extended version E-LE incorporating different types of meta-information (II-VI). P-values for the difference between methods I-VI and method I (p_1) and method VI (p_2) are presented. † stands for p<0.001 The results for method V are significantly different from all other results with p<0.001 apart from method VI. The bottom rows of the table present classification rates when using different types of metadata only.

	AD vs CN			P-MCI vs S-MCI			P-MCI vs CN			p_1	p_2
	ACC	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE		
I: LE	84	81	88	62	50	72	80	74	85	n.a.	†/†/†
II: E-LE with ApoE	84	81	87	63	51	73	81	75	85	0.61/†/†	†/†/†
III: E-LE with Aβ_{42}	86	83	90	65	59	69	83	78	87	†/†/†	0.68/†/†
IV: E-LE with hippo. vol.	87	84	90	65	56	74	83	80	86	†/†/†	0.02/†/†
V: E-LE with Aβ_{42}, hippo. vol.	88	85	90	68	65	70	85	82	88	†/†/†	0.03/0.06/0.06
VI: E-LE with Aβ_{42}, hippo. vol., ApoE	88	85	90	67	64	70	86	81	89	†/†/†	n.a.
ApoE only	67	59	75	57	52	64	68	63	75	†/†/†	†/†/†
Aβ_{42} only	75	64	85	64	51	81	73	63	84	†/†/†	†/†/†
Hippo. vol. only	74	74	74	61	58	62	72	73	70	†/†/†	†/†/†

Table 6.3: Statistics from regressing MMSE versus d manifold coordinates using a multiple linear model. An improvement of statistics can be observed when incorporating metadata into the manifold learning process. Results are presented for d=15/d=1

	t	residual	R^2	F	p
I	10.6/10.1	1.76/1.91	0.29/0.20	11.1/102.7	†
II	-10.7/-10.2	1.76/1.91	0.29/0.20	11.3/104.2	†
III	-11.0/-10.4	1.73/1.89	0.32/0.21	12.5/109.7	†
IV	11.1/10.5	1.73/1.89	0.30/0.21	11.8/109.7	†
V	-11.6/-10.9	1.71/1.87	0.33/0.22	13.6/119.5	†
VI	11.9/11.4	1.73/1.84	0.32/0.24	12.5/131.2	†

Table 6.4: Classification accuracy (ACC), sensitivity (SEN) and specificity (SPE) when incorporating metadata for classification into the SVM featurevector. p<0.001 for differences with the according methods (III, IV) in Table 6.2 are labeled with a **bold** classification accuracy.

	AD vs CN			P-MCI vs S-MCI			P-MCI vs CN		
	ACC	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE
ApoE	84.7	79.9	89.0	63.2	52.7	71.6	81.1	76.2	84.8
Aβ_{42}	86.3	83.2	89.1	64.6	57.9	69.6	**82.3**	79.3	85.2
hippo. vol.	**86.0**	81.6	89.3	**63.7**	52.3	71.9	82.9	79.0	86.5

Learning a low-dimensional manifold from a combined similarity measure

The similarity between two instances $z_i, z_j \in Z_c$ of a continuous variable can be defined as $\hat{s}_{ij} = \frac{\text{abs}(z_i - z_j)}{\max(Z_c)}$. With the intensity-based similarity s_{ij}, the edge weight used for manifold learning with LE (Equations 6.1-6.4) can then be defined using a combined similarity measure:

$$w_{ij} = \begin{cases} s_{ij} + \alpha \hat{s}_{ij}, & \text{if } i \in \mathcal{N}_i \text{ or } j \in \mathcal{N}_j \\ 0, & \text{else.} \end{cases} \tag{6.15}$$

where $\mathcal{N}_x$ describes the k nearest neighbors to subject x and α defines the relative influence of the two similarity measures. The classification performance of this approach to find an LE embedding for $\hat{s}_{ij}$ defined by hippocampal volume and Aβ_{42} concentration was evaluated. Varying $\alpha \in [0, 7]$ and applying SVM-based classification on the resulting manifold coordinates with the procedure described in Section 6.3.5, results in the classification rates displayed in Figure 6.6.

6.4 Discussion

In this chapter, a method to extract biomarkers from MR brain images, combining imaging information with non-imaging metadata was presented. Laplacian eigenmaps were used to derive a nonlinear and low-dimensional representation of a set of images. The graph defined by pairwise similarities and used to define the Laplacian eigenmaps objective function is extended by *support nodes* representing metadata. Weights defined from all image nodes to all support nodes incorporate the metadata into an extended objective function. Optimizing this target function leads to an embedding that is expressed by both pairwise image similarities and the similarity represented by the metadata. The proposed method was evaluated on a large and diverse clinical dataset (ADNI). The presented results show that the proposed method is able to produce a classification accuracy between clinical groups with an accuracy that compares favorably to established and state-of-the-art methods in neuroimaging. Cuingnet et al. [38] recently presented the comparison of ten different methods for classification on a subset of ADNI similar to that used in this study. These methods comprise five high dimensional voxel-based approaches, three methods based on cortical thickness and two methods based on the hippocampus. Using only imaging similarities, the proposed manifold-based method outperforms the majority of the ten methods in individual classification experiments and lags behind only slightly to the STAND-score [144] when averaging results over the three clinical pairings evaluated. Compared to most of the other nine methods, the STAND-score had a relatively good performance in the identification of progressive MCI subjects, resulting in classification accuracies of 80%, 71% and 81% for AD vs CN, S-MCI vs P-MCI, P-MCI vs CN respectively.

Incorporating non-imaging information into the manifold learning step, yields substantial and significant improvements in classification accuracy. The individual use of ApoE genotype, the concentration of Aβ_{42} and hippocampal volume, improves classification rates. Using all metadata in one step, further improves results to 88% for AD vs CN, 67% for P-MCI vs S-MCI and 86% for P-MCI vs CN. These results highlight the potential role of such metadata as suitably complementary information to MR

image data in future studies. In addition to classification performance, the ability of the learned manifold to predict clinical variables was evaluated. Learning a multiple linear regression model of MMSE versus manifold coordinates, leads to significantly improved results compared to what has been published using similar data. Gerber et al. [63] report $R^2 = 0.05$ and a residual of 2.37 when regressing MMSE versus the first manifold coordinate. Incorporating metadata led to further improved regression statistics in this sample dataset.

Two alternative approaches to incorporate non-imaging information into a manifold classification setting were discussed. The proposed method shows better classification accuracy compared with a method in which image similarities and non-imaging similarities are combined before performing manifold learning. However, tuning the weighting factor between the concatenated similarities on the test images, did lead to results comparable with the proposed method for the sub-comparisons of AD vs CN and P-MCI vs CN. Compared to an approach where manifold features are combined with a meta-variable before performing SVM-based classification, the proposed method gave slightly superior performance. The strength of the proposed method, however, lies in the unified representation of information taken from different measurements. This enables not only classification but can also help in visualizing the determined biomarker in a clinical environment. Plots of the form shown in Figure 6.4 can potentially enhance interpretation of computer-aided diagnosis (CAD) systems, such as the one developed in PredictAD (Chapter 1.2). A clinician can locate the patient studied relative to all other database cases providing information about the severity of the disease not only the on/off-classification result. Furthermore, the capability to define a single continuous biomarker facilitates the definition of regression models such as the one presented in Section 6.3.6. Tests were carried out to evaluate the influence of the number of embedding dimensions m on training data. Robust results where achieved for $m \in [6, 15]$. Assuming normalized weights defined on the metadata and a normalized pairwise similarity measure, the weighting factor γ that dictates the influence of metadata on the manifold coordinates, can be set globally.

While individually tuning γ for every type of metadata is expected to lead to better results, a weighting based on training data was determined and used for all experiments in order to work with a more realistic setting. Many state-of-the-art methods for the extraction of biomarkers for AD from MR images are computationally expensive (run-time of hours to days) or require complex a-priori information (e.g. manual segmentation in atlas-based methods) [38]. The proposed method provides a fast and robust alternative to classify subjects that is generic and data-driven. The computational time to classify a new subject is around 10 minutes on a standard 8-core desktop machine including registration to a template space and feature extraction (measuring pairwise similarities to a training set) as well as classification with SVMs.

In the next chapter, an extension to the developed framework is presented, describing different ways to incorporate longitudinal information into the manifold learning process.

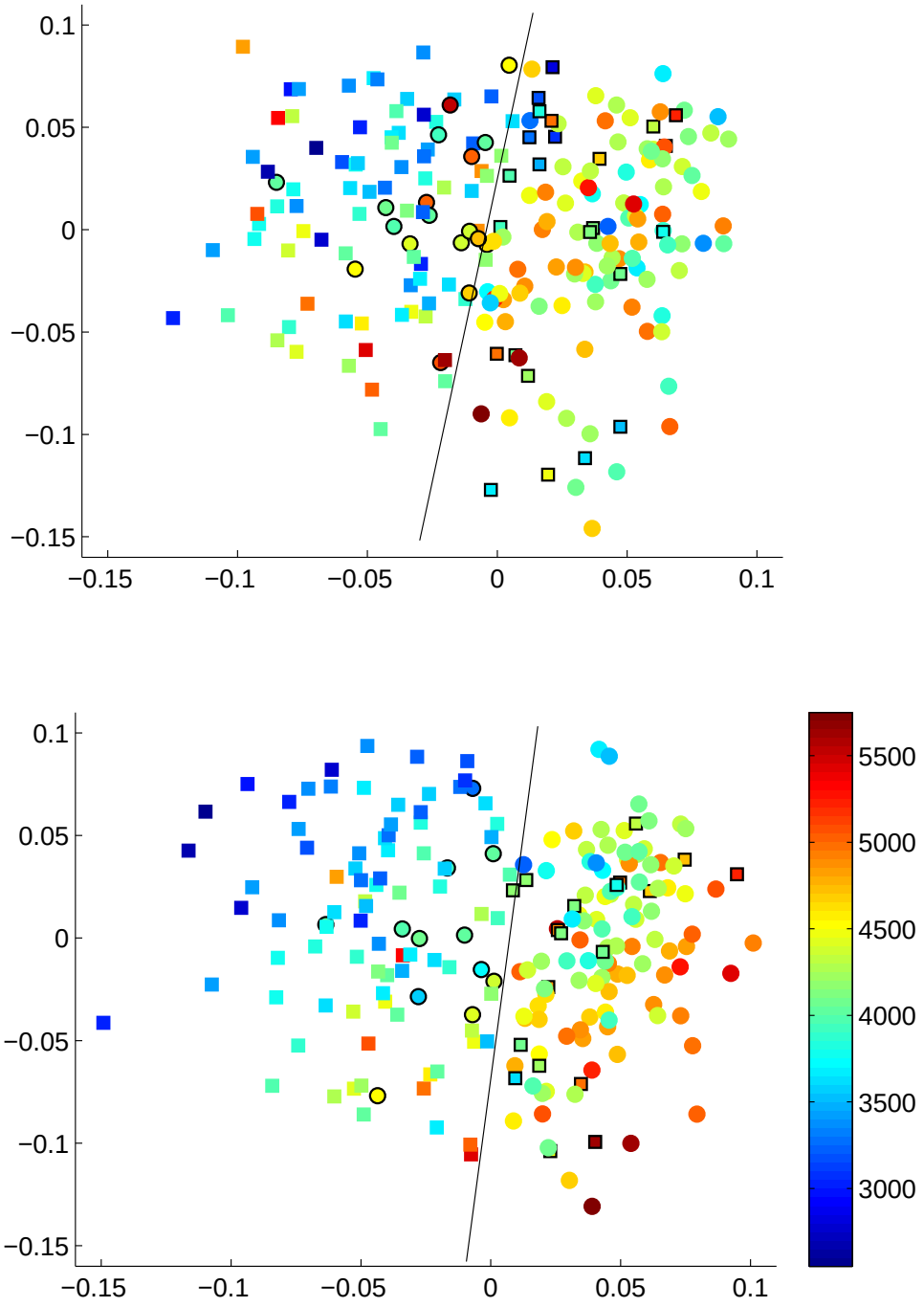
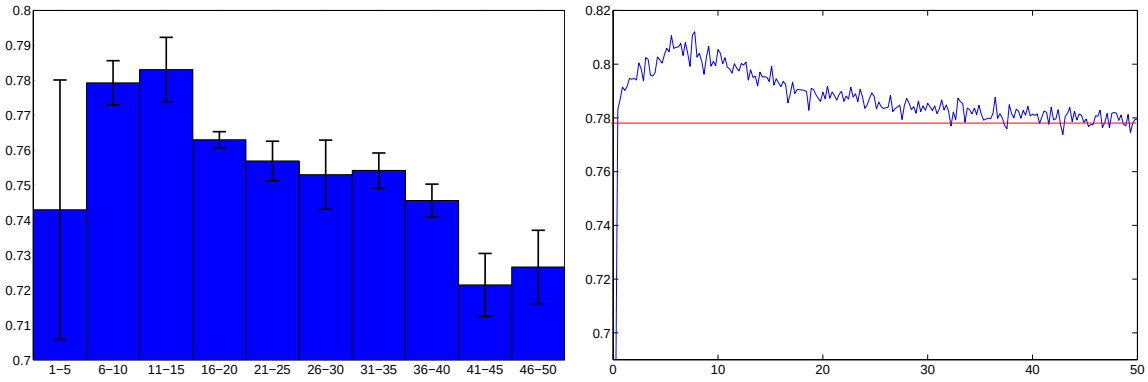


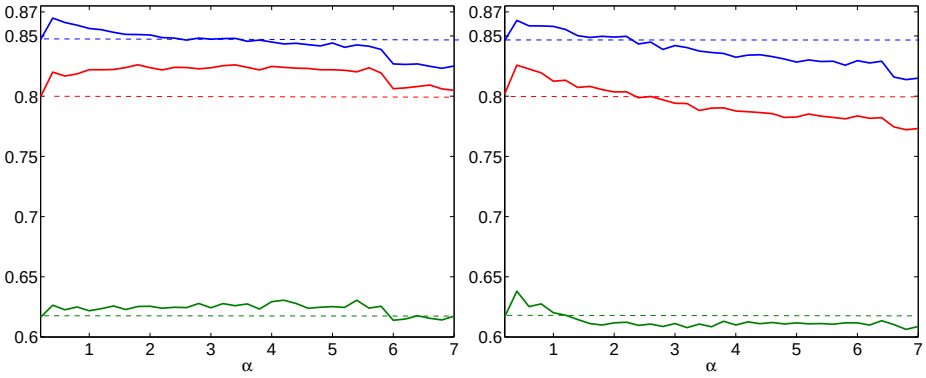
Figure 6.4: Standard embedding using Laplacian eigenmaps based on pairwise image similarities only (top). Extended embedding using the proposed method with hippocampal volume as metadata (bottom). 103 AD patients are represented by squares, 116 healthy controls by circles. Hippocampal volume (cm^3) is encoded in the marker color. A SVM separating hyperplane in 2 dimensions is displayed. Misclassified subjects with both approaches are highlighted by a black outline (42 with LE, 31 with E-LE). An improved separating ability can be observed in the extended embedding especially for subjects close to the separating plane in the original embedding.



(a) Dimension d

(b) Weighting factor γ

Figure 6.5: (a): Classification rate in the training data set when varying the dimension of the low-dimensional manifold between 1 and 50. Results are presented as mean rates over ten bins covering five dimensions each. The high standard deviation observed in the first bin results from a very low classification rate with d=1. (b): Classification rate when varying γ between 1 and 50 evaluated for $d \in [6, 15]$.



(a) Hippocampal volume

(b) Aβ_{42}

Figure 6.6: Classification accuracy obtained from defining a combined similarity measure incorporating both imaging and non-imaging information before performing manifold learning. AD vs CN: blue; S-MCI vs P-MCI: green; CN vs P-MCI: red. Results with hippocampal volume and Aβ_{42} are presented over an increasing influence of the metadata. The dotted lines indicate the classification accuracy obtained with image similarities only.

Chapter 7

Manifold learning incorporating longitudinal data

This chapter is based on:

Robin Wolz, Paul Aljabar, Joseph V. Hajnal, Daniel Rueckert. "Manifold learning for biomarker discovery in MR imaging". *Workshop on Machine Learning in Medical Imaging, MICCAI 2010*, Beijing, China, September 2010

Abstract

This chapter presents an extension of the manifold classification framework described before. Here, a low-dimensional manifold is described by both the inter- and intra-subject variation in brain MR image data. The key contribution is the incorporation of longitudinal image information in the learned manifold. In particular, simultaneously embedding baseline and follow-up scans into a single manifold is compared with the combination of separate manifold representations for inter-subject and intra-subject variation. The proposed methods are applied to 362 ADNI subjects to classify healthy controls, subjects with AD and subjects with MCI. Learning manifolds based on both the appearance and temporal change of the hippocampus, leads to correct classification rates comparable with those provided by state-of-the-art automatic segmentation estimates of hippocampal volume and atrophy.

7.1 Introduction

Chapter 6 described a method to learn a low-dimensional manifold based on inter-subject brain variation and subject meta-information representing cross-sectional differences across the population. The typical patterns of change in the aging brain are altered by neurodegenerative diseases such as AD. This makes structural change over time a reliable biomarker, e.g., [57]. To characterize brain development, this chapter considers longitudinal brain studies are considered where MR scans at baseline and after different follow-up intervals are available. The inspection of scans from a single timepoint allows inferences about the *inter-subject* variation in the study population. Comparing two scans taken from the same subject at different timepoints, yields insights into *intra-subject* variation. Many researchers in computer vision have addressed the problem of embedding images while considering both intra- and inter-subject variation, e.g., [133, 23]. The conclusion is that separating intra- and inter-subject variation can lead to a more powerful model. To further investigate such aspects, two approaches are proposed to model longitudinal variation. In the first approach, follow-up scans are simultaneously embedded together with their baseline images. In the second approach, a separate manifold is learned based on the difference images between two timepoints representing intra-subject variation.

In the evaluation, the 362 ADNI subjects are used for which at least three timepoints (baseline, month 12 and month 24) were available at time of retrieval (February 2010).

7.2 Method

7.2.1 Manifold learning for cross-sectional data

As described in Chapters 2 and 6, in manifold learning a set of high-dimensional images $\mathbf{X} = \{\mathbf{x}_1, ..., \mathbf{x}_N\} \in \mathbb{R}^D$ is represented in a low dimensional space as $\mathbf{Y} = \{\mathbf{y}_1, ..., \mathbf{y}_N\} \in \mathbb{R}^d$ with $d \ll D$. As before, Laplacian eigenmaps (LE, Chapter 2.4.2)

are used to perform the embedding to be able to directly take image similarities as an input measure. In order to learn a low dimensional representation of cross-sectional data, cross correlation in a region around hippocampus and amygdala (see also Chapter 6.3.2) is evaluated in a k-nn neighborhood of the input data to define the weights w_{ij} in the objective function of Laplacian eigenmaps:

$$\phi(\mathbf{Y}) = \sum_{ij} \|\mathbf{y}_i - \mathbf{y}_j\|^2 w_{ij} = 2\mathbf{Y}^T\mathbf{L}\mathbf{Y} \tag{7.1}$$

7.2.2 Manifold learning for longitudinal data

One natural approach to account for longitudinal information in the presented manifold learning framework, is to apply LE to a set of images consisting of both baseline- and follow-up scans. It has been shown, however, that inter-subject variation can dominate the embedding and the relatively subtle intra-subject variation can be lost in the low-dimensional manifold [133, 23]. To further investigate this, two different approaches are proposed to incorporate longitudinal information into the classification framework: (a) embedding both timepoints separately and (b) independently embedding baseline images and difference images representing longitudinal change.

The set of images in a longitudinal study with M visits can be defined as $\mathbf{X}_{ij} = \{x_{ij} : 1 \leq i \leq N, 0 \leq j \leq M-1\}$ where N is the number of subjects.

The images acquired at the J-th follow-up visit, $\mathbf{X}_J = \{\mathbf{x}_{iJ} : 1 \leq i \leq N\}$, are rigidly aligned with the according baseline scans $\mathbf{X}_0 = \{\mathbf{x}_{i0} : 1 \leq i \leq N\}$ and resampled in the baseline coordinate system. A set of difference images $\mathbf{X}_{\Delta J} = \{\mathbf{x}_{i\Delta J} : 1 \leq i \leq N\}$ is then derived with $\mathbf{x}_{i\Delta J} = \mathbf{x}_{iJ} - \mathbf{x}_{i0}$.

During LE, the weights w_{ij} representing the similarity of two images $\mathbf{x}_i$ and $\mathbf{x}_j$ determine the coordinate embedding produced through the objective function in Equation 7.1. To inspect different approaches for longitudinal modeling, the weights matrix $\mathbf{W}$ is constructed from different sets of data. With the superscript S denoting the data set used to construct $\mathbf{W}$, the LE mapping for scan j of subject i is given by $\mathbf{x}_{ij} \rightarrow \mathbf{y}_{ij}^S$ for scans and $\mathbf{x}_{i\Delta j} \rightarrow \mathbf{y}_{i\Delta j}^S$ for difference images. With these definitions, the inter-subject

variation at baseline is defined by $\mathbf{y}_{i0}^{X_0}$. The combined coordinate embedding $\mathbf{y}_{ij}^{X_0 \cup X_j}$ is learned from both variation at baseline and intra-subject change at timepoint j, and $\mathbf{y}_{i\Delta j}^{X_{\Delta j}}$ finally captures longitudinal change only. Three different feature vectors are defined from the above embeddings, two of which are obtained by concatenating scans' embedding coordinates:

A Baseline scans in one manifold: $\mathbf{y}_{i,A} = \left(\mathbf{y}_{i0}^{X_0}\right) \in \mathcal{R}^d$

B Two scans per subject in one manifold: $\mathbf{y}_{ij,B} = \left(\mathbf{y}_{i0}^{X_0 \cup X_j}, \mathbf{y}_{ij}^{X_0 \cup X_j}\right) \in \mathcal{R}^{2d}$

C Baseline / difference images in two manifolds: $\mathbf{y}_{ij,C} = \left(\mathbf{y}_{i0}^{X_0}, \mathbf{y}_{i\Delta j}^{X_{\Delta J}}\right) \in \mathcal{R}^{2d}$

7.3 Experiments and results

7.3.1 Subjects

The proposed method was applied to 362 subjects from the ADNI study consisting of patients with mild AD (N=83, mean MMSE 23), MCI (N=165, mean MMSE 27) and healthy control subjects (CN, N=114, mean MMSE 29). For each subject, T1-weighted 1.5T MR images were available for the baseline, 12 month and 24 month scans. For the MCI group, 75 subjects were diagnosed with AD after baseline scanning. Progressive (P-MCI) and stable (S-MCI) groups were therefore analyzed independently. For eight subjects in the MCI group and two subjects in the AD group, a reversion to CN and MCI respectively was reported and these subjects were excluded from the analysis.

7.3.2 Parameter settings

The optimal neighborhood size, k, for the graphs used to learn the embeddings depends on the dataset. Following the findings for ADNI data presented in Chapter 6, the parameter was set to $k = 20$. There is no defined procedure to establish the best dimension in a learned manifold with LE. Following the results presented in Chapter 6, average classification results when evaluating the framework for $d \in [6, 15]$ are reported.

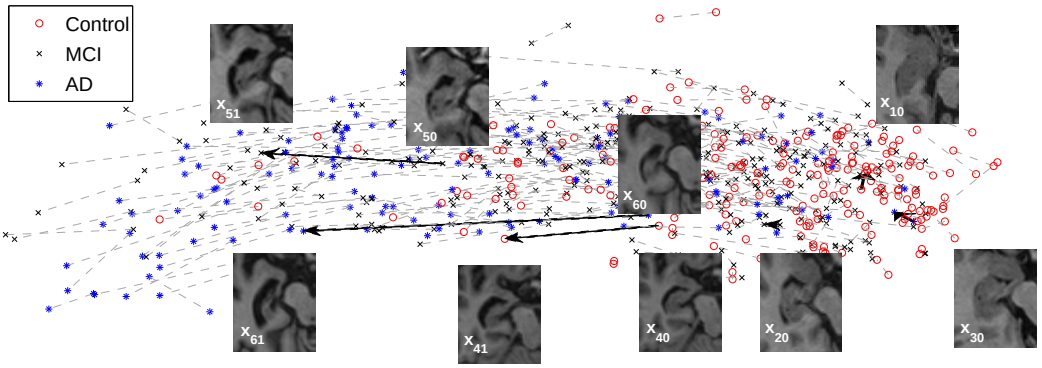
Table 7.1: Correct classification results in percentages using different feature vectors based on scans' coordinates in the learned manifolds (Section 7.2.2). Vector A is based on baseline features only. For vector B, baseline and follow-up scans (after 12 or 24 months) are together embedded in one manifold. Vector C consists of features taken from the baseline embedding and a separate embedding of longitudinal image differences. Average classification rates (ACC), sensitivity (SEN) and specificity (SPE) are displayed when varying the dimension of the manifold $l \in [6, 15]$.

Feature	AD vs CN			P-MCI vs S-MCI			P-MCI vs CN		
	ACC	SEN	SPE	ACC	SEN	SPE	ACC	SEN	SPE
A: $\mathbf{y}_{i,A}$	84.2	78.6	88.3	62.0	58.8	64.7	80.7	72.4	86.2
B: $\mathbf{y}_{i1,B}$	84.9	79.8	88.6	64.0	61.0	66.6	80.6	73.4	85.4
B: $\mathbf{y}_{i2,B}$	87.9	84.9	90.0	62.1	61.1	63.0	81.9	76.1	85.7
C: $\mathbf{y}_{i1,C}$	83.7	78.6	87.3	65.9	60.7	70.2	77.9	68.9	83.9
C: $\mathbf{y}_{i2,C}$	85.4	82.3	87.7	67.3	63.7	70.3	82.9	75.9	87.5

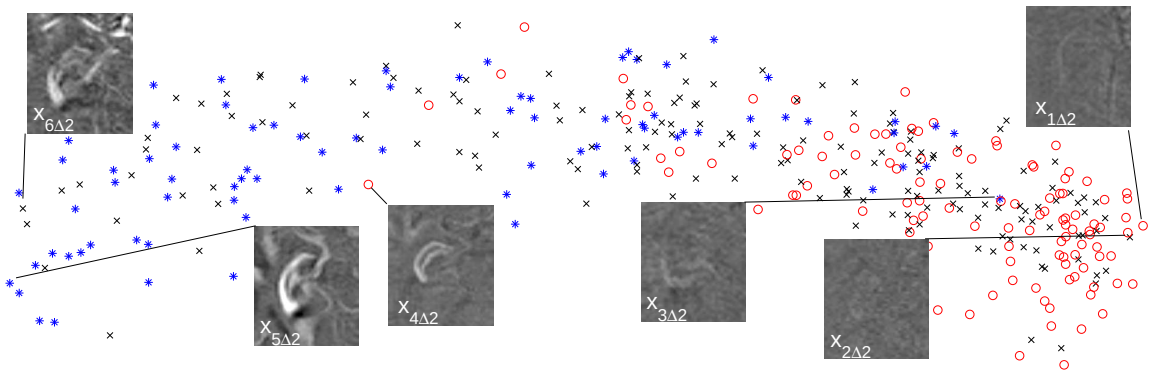
7.3.3 Classification

All 1086 study images were aligned with a coarse non-rigid registration [118] to the MNI152 brain template. Follow-up images after 12 and 24 months were rigidly aligned with their baseline scans to derive difference images. These images were aligned with the brain template using the deformation field estimated for the baseline scan. Pairwise similarities were evaluated between all brain images and between the sets of difference images representing change over a given time period (month 12 / month 24). Similarities were evaluated over the region around hippocampus and amygdala that has also been used in Chapter 6 to measure pairwise similarities.

Following the approach described in Chapter 6.2.2, Linear SVMs were then used to define a separating hyperplane between two subject groups based on the feature vectors $\mathbf{y}_{i,A}, \mathbf{y}_{ij,B}, \mathbf{y}_{ij,C}$. A leave-25%-out approach was applied: for each repetition, 75% of the subjects in both groups were randomly selected and used to train a SVM classifier. The remaining 25% of subjects in both groups were used as a test set. 1000 repetitions were applied for all pairings of clinically interesting groups. Classification rates for the three feature sets are displayed in Table 7.1. A visualization of the 2D-embedding for both longitudinal methods is given in Figure 7.1 where the follow-up images used were the 24 month scans ($j = 2$). In Figure 7.1 (a), both, baseline images $\mathbf{X}_0$ and 24 month follow-up images $\mathbf{X}_2$, are embedded together. Figure 7.1 (b) shows the results



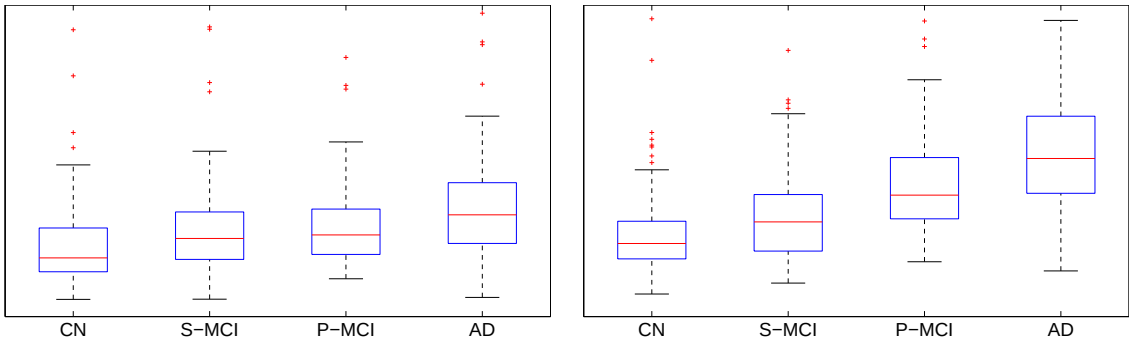
(a) Simultaneous embedding of baseline and 24 month follow-up scans. Trajectories are displayed for each subject as a dashed line. Subjects with highlighted trajectories are also illustrated in (b) below. Where changes are very small, only baseline images are displayed for reasons of space.



(b) Embedding of difference images between baseline and 24 month scan ($X_{\Delta 2}$). For each clinical group, subjects with extremely high and low values for the first embedding coordinate are displayed.

Figure 7.1: 2D visualizations of manifolds incorporating longitudinal information. Exemplar images are labeled $\mathbf{x_{ij}}$ and $\mathbf{x_{i\Delta j}}$ with $i = 1, ..., 6$ and $j = 0, 2$ where i represents the subject id and j the visit number.

of embedding the difference images $\mathbf{X}_{\Delta 2}$ representing longitudinal change in a separate manifold. Exemplar images of the six subjects that lie at extreme positions within each group in the difference embedding (b), are displayed in both manifolds. It can be seen that the extremes of the longitudinal changes, large and small, displayed by the difference images are also well represented in embedding (a) resulting in extremely long and short trajectories between the timepoints respectively. Figure 7.2 shows the box-and-whisker plots for the distance a subject "moves" in the combined manifold over 12 and 24 months. While there is only a slight trend of a difference in the movement over 12 months, a clearer separation between the clinical groups can be observed in the movement over 24 months.



(a) Month 12 (b) Month 24

Figure 7.2: Box-and-whisker plots for the distance between longitudinal images in a manifold learned from images at several timepoints.

For comparison, Table 7.2 shows classification results based on automatically determined hippocampal baseline volumes as described in Chapter 4 and atrophy rates as described in Chapter 5 for the subset of images used in this study. Additionally, canonical correlation analysis was applied to measure the correlation of features in the defined d-dimensional manifolds with hippocampal volume and atrophy rates. The correlation coefficient r between baseline volume and the coordinates in the baseline embedding $\mathbf{y}_{i0}^{X_0}$ is reported. In addition, the correlation between the vector $\mathbf{y}_{iJ}^{X_0 \cup X_j} - \mathbf{y}_{i0}^{X_0 \cup X_j}$ describing the trajectory between two subjects in a combined embedding (see Figure7.1 (a)) and atrophy is presented. Finally, the correlation of atrophy with the coordinates $\mathbf{y}_{i\Delta J}^{X_{\Delta J}}$ in the difference embedding (see Figure 7.1 (b)) is presented.

Table 7.2: Classification results based on hippocampal baseline volume (Chapter 4) and atrophy (Chapter 5) over 12 and 24 months. The second part of the table shows the correlation of coordinates in the learned manifolds with baseline volume and atrophy. $d = 20$ coordinates of $\mathbf{y}_{i0}^{X_0}$, $\mathbf{y}_{i\Delta J}^{X_{\Delta J}}$ and $\mathbf{y}_{iJ}^{X_0 \cup X_j}$ are used to determine r. a: $p < 10^{-4}$

	AD vs CN	P-MCI vs S-MCI	P-MCI vs CN	r for $\mathbf{y}_{i0}^{X_0}$	r for $\mathbf{y}_{iJ}^{X_0 \cup X_j} - \mathbf{y}_{i0}^{X_0 \cup X_j}$	r for $\mathbf{y}_{i\Delta J}^{X_{\Delta J}}$
Baseline vol.	75%	59%	73%	0.62^a	-	-
Atrophy M12	82%	66%	76%	-	0.63^a	0.75^a
Atrophy M24	86%	67%	83%	-	0.73^a	0.87^a

7.4 Discussion and conclusion

This chapter presented an extension of the manifold classification approach described in Chapter 6, extending it from the use of inter-subject appearance in a data set at baseline to also incorporating intra-subject changes over time. While classification rates based on baseline appearance on the used dataset are in line to the baseline results presented in Chapter 6, a significant improvement can be achieved when considering longitudinal information. Using a longitudinal embedding based on difference images, the classification accuracy between P-MCI and S-MCI subjects can be increased from 62% to 67% with 24 month follow-up scans. A combined embedding of baseline and follow-up scans on the other hand allows to substantially improve AD vs CN classification. A combination of both approaches is potentially able to achieve a more stable improvement of classification accuracy. Furthermore, incorporating metadata at baseline and follow-up as described in Chapter 6 may lead to a more accurate representation of the population and hence a more accurate biomarker.

The presented results show that the application of the proposed framework to similarities based on a region of interest (ROI) around hippocampus and amygdala leads to classification results comparable if not superior to those obtained from automatically determined hippocampal volume and atrophy. This shows that the information that may be learned about a subject's clinical state from estimates of hippocampal volume and atrophy is also encoded in the manifolds learned from inter- and intra-subject variation in the ROI respectively. These conclusions are also supported by the significant correlation that were found between hippocampus volume and atrophy with manifold coordinates.

The results presented in this chapter are revised from work published in Wolz et al. 2010 [150]. Following insights from the work presented in Chapter 6 in this thesis, data processing has been optimized. The region of interest has been extended from covering only the hippocampus to also including amygdala. After a more detailed evaluation, the used manifold dimensions have been generally restricted to $d \in [6, 15]$ from $d \in [1, 20]$. Furthermore, manifold coordinates are corrected for subject age

using a multiple linear regression model. These factors led to significantly improved classification rates compared to the original publication.

Chapter 8

Comprehensive analysis of MR-derived biomarkers

This chapter is based on:

Robin Wolz*, Valtteri Julkunen*, Juha Koikkalainen, Eini Niskanen, Dong Ping Zhang, Jussi Mattila, Daniel Rueckert, Hilkka Soininen, Jyrki Lötjönen. Comprehensive Analysis of MRI Images in Early Diagnostics of Alzheimers Disease. *Submitted*, 2011

Abstract

Using the different biomarkers proposed in this thesis, this chapter aims to assess the improvement in classification rate that can be achieved by combining features from different structural MRI analysis techniques. Classification into the diagnostic groups was done with automatic MRI methods including hippocampal volume and atrophy, cortical thickness, tensor-based morphometry and manifold-based learning. The results show that a comprehensive analysis of MRI images combining multiple methods improves classification accuracy and predictive power in detecting early AD. The increase in classification accuracy obtained with repeated follow-up MRI may not justify the additional cost and waiting time.

*Both authors contributed equally

8.1 Introduction

Recent studies focusing on structural MRI methods have reached correct classification rates (ACC) of 76-94 % in identifying healthy controls (CN) from patients with AD and 64-82 % in predicting which MCI subjects will convert to AD in the imminent future [29, 62, 98, 99, 107, 114, 41]. However, comparison of the results is not straightforward since the study populations and classification methods differ substantially. Also not all published results are validated by using separate training/testing sets or cross-validation [98, 99, 41], which can lead to overestimation of a method's accuracy and compromise the generalizability of the results.

It has been shown that the early diagnostics of AD can be improved by using multiple different biomarkers simultaneously. Like the results presented in Chapter 6, most of these studies have combined MRI-based markers with biomarkers based on positron emission tomography (PET) [74], cerebrospinal fluid (CSF) [39, 46] or both [89, 91, 146], but the results vary from no additional benefit [39, 89] to significant improvement [74, 46]. However, availability of all three biomarkers (CSF, PET, MRI) is not very common in clinical practice. Obtaining all measures is also laborious for the patient and clinician, induces delays and increases the costs of the diagnosis significantly.

Performance of different structural MRI methods have been recently compared [38], but the full potential of structural MRI has not been investigated thoroughly. It is not clear (I) which structural MRI methods provide best results, (II) if the use of several structural methods simultaneously provides an improvement or (III) if the classification accuracy and predictive power can be enhanced by assessment of repeated MRI scans during follow-up. In order to find answers to these critical questions this chapter combines the biomarkers presented in this thesis with other fully automatically extracted, state-of-the-art MRI based features for AD. In addition to hippocampal volume (HV), hippocampal atrophy (HA) and manifold-based learning (MBL), tensor-based morphometry (TBM) as well as cortical thickness (CTH) are

combined to perform an overall analysis of classification accuracy.

Using HV, HA, TBM and MBL, experiments were carried out on Baseline, Month 12 and Month 24 images taken from 477 subjects from the ADNI database. MRI features extracted from these images were used separately and combined to perform classification between CN and AD, to predict a conversion from MCI to AD (classification of stable MCI (S-MCI) from progressive MCI (P-MCI)) and to detect prodromal AD (CN/P-MCI classification). In a separate experiment, HV, TBM, MBL and CTH were combined on a separate image set consisting of 364 ADNI baseline scans. A linear discriminant analysis (LDA) was performed to combine the features obtained from the individual methods.

8.2 Materials and Methods

8.2.1 Subjects

All ADNI subjects (152 CN, 112 S-MCI, 110 P-MCI, 103 AD) for which a 1.5T T1-weighted MRI scan at baseline, month 12 and month 24 was available in September 2010 were included in the analysis presented in this chapter. All subjects for which a reversion from AD to MCI (N=4) or from MCI to CN (N=21) has been reported so far were excluded from the study.

The toolbox applied for cortical thickness measurement did not achieve satisfactory results on all 477 study subjects. The combination of all available features was therefore evaluated on the independent subset of 364 baseline images for which a CTH analysis was acceptably performed. Figure 8.1 gives an overview on the inclusion / exclusion criteria as well as the different classification tasks performed.

The results presented in this chapter are based on joined work carried out in PredictAD. HV calculation was performed by JL. HA and MBL analysis was carried out by RW. TBM results are obtained by JK and VJ performed the CTH measurement. The combined analysis was carried out by JK.

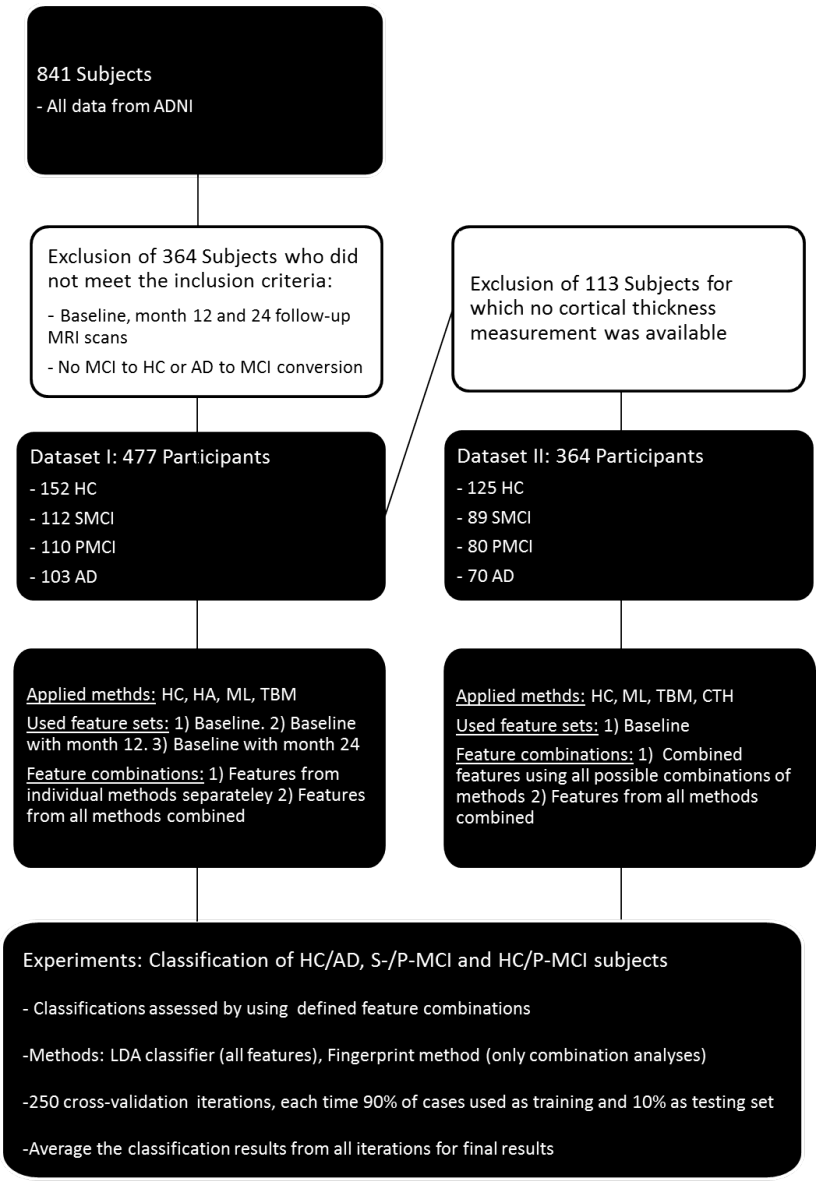


Figure 8.1: Inclusion / exclusion criteria

Hippocampal volume

Baseline hippocampal volume was measured using an approach based on fast and robust multi-atlas segmentation [104]. In this approach, multi-atlas label propagation is applied in combination with atlas selection to obtain the hippocampus segmentation. A set of hippocampus atlases is selected from a pool of atlas images according to image similarity with the query image. After registering all atlases to the query image, a spatial prior is generated from the multiple label maps. This spatial prior is then used to obtain a final segmentation based on an expectation maximization (EM)

segmentation algorithm.

Hippocampal atrophy

Hippocampal atrophy over 12 and 24 months was measured using the method for a simultaneous and consistent segmentation described in Chapter 5 of this thesis.

Cortical thickness

CTH is measured in the baseline T1-weighted structural MR images by using an automated computational surface-based method [95, 87]. The pipeline includes registration of the images to a standard space, correction of error caused by intensity non-uniformities, tissue segmentation, partial volume effect (PVE) magnitude estimation, creation of two polygon meshes on the cortical surfaces and calculation of the distance between the adjacent nodes on the surfaces using the t-link metric. As a result the pipeline measures CTH at sub-millimeter accuracy in 40962 nodes per hemisphere.

Tensor-based morphometry

In tensor-based morphometry, TBM (or deformation-based morphometry, DBM), features are extracted from the deformation field obtained from registering a set of subjects to a template space (see Chapter 2). The TBM analysis used here was performed using a recently presented multi-template approach [94, 21]. Instead of using just one template to which all the study images are registered, 30 randomly selected images from the ADNI database were used as template images. Each study image was registered to each template image. To combine the results of each template image, the template images were registered to the mean anatomical template generated from the 30 template images, and all the results were normalized to and presented in this reference space. For the classification, the mean Jacobian was computed in the 83 ROIs defined in the Hammers brain atlas (Chapter 2.1.1) for each study image and each template. The mean was computed only in atrophic voxels and it was weighted based on the voxel-wise group-level p-value. The feature values for the classification

were obtained by averaging the ROI-wise mean Jacobians of all the 30 templates. The atrophic voxels and the p-values were obtained from voxel-wise t-tests computed using a separate training set that was not used in the evaluation of classification accuracy.

Manifold-based learning

Based on the methods described in Chapters 6 and 7, manifold-based features were extracted. All three timepoints of every subject were used to find a single manifold-embedding as described in Chapter 7. Following the results presented in Chapter 6, the first 15 embedding coordinates for each subject and timepoint are then used as a feature.

8.2.2 Statistical analysis

Statistical Regions-of-Interest

In the analysis of CTH and TBM, information on the regions with statistically significant group-level differences between two study groups was used to determine the feature values. This information was computed from the baseline images of those ADNI cases for which month 12 and/or month 24 follow-up images were not available (N = 295 for TBM and N = 233 for CTH), and hence were not used to evaluate the classification performance.

In CTH, the feature values were computed only from the regions with statistically significant differences between the two study groups.

To examine statistical differences in CTH between the study groups a t-test was performed in every cortical node in both hemispheres using Matlab. A correction for multiple comparisons was done using the false discovery rate (FDR)-correction method [61]. Age and gender were used as nuisance variables in all CTH analyses. The level of significance was set to $p = 0.05$. In order to find the areas with probable disease related cortical thinning the nodes with the lowest absolute t values were discarded. Limitations used to form the CTH ROIs from the statistical analysis in the MCI classifications were: $t_{min} = 3$ and number(nodes) > 100 / ROI. This resulted in

separate ROIs representing the most significant difference between groups of interest. The mean CTH values calculated separately for each ROI as well as for all significant points together were then taken as CTH features in the classification tasks. In the TBM analysis, feature computation was constrained to atrophic voxels as it is known that AD causes atrophy in cerebral cortex and sub-cortical structures. Also, weighting was used to emphasize the regions with statistically highly significant group-wise differences. The required information was extracted from the data using t-tests.

8.2.3 Classification

All feature values were corrected for age and gender using a linear regression model where control subjects were used as the training set, i.e., the normal, not disease-related, age and gender related differences in the classification features were removed. Feature selection was carried out on the corrected feature sets using stepwise regression [44]. A leave-10%-out strategy was applied where 90% of the subjects were randomly selected and used to train classification parameters and the remaining 10% were classified accordingly. The reported classification results are averaged over 250 iterations of this procedure. The classification procedure is illustrated in Figure 8.1.

Linear discriminant analysis (LDA)

Linear discriminant analysis (LDA) was used to perform classification based on the defined feature sets. LDA is a widely used technique to find a linear combination of features to best separate several classes [90]. In this work, LDA was used as implemented in the *classify* function in Matlab with a multivariate normal density model with uninformative priors (p=0.5).

8.3 Experiments and results

8.3.1 Image sets

Baseline characteristics of the full data set, referred to as dataset I, for which scans at baseline, month 12 and month 24 were available are presented in Table 8.1. There were differences between the study groups in all variables except age ($p < 0.05$). There were more men than women in all groups besides the AD group. MMSE scores were significantly different in the pairwise comparisons between all study groups. Compared to controls, all other groups had significantly shorter education. Carriers of the APOE4 allele were substantially more abundant in the P-MCI and AD groups.

The second image set (dataset II) used in this study is the subset of baseline scans used in dataset I for which a cortical thickness measurement was available. The 364 subjects consist of the following sub-groups: 125 CN, 89 S-MCI, 80 P-MCI, 70 AD. Demographic and clinical data for dataset II does not differ significantly (at threshold p=0.05) from the full subset, dataset I, described above when comparing whole sets and when comparing individual clinical groups.

8.3.2 Classification results using dataset I

CN vs AD classification

The classification results of CN and AD subjects are presented in Table 8.2. The results for ACC / SEN / SPE lie in the range of 80-89%. The manifold based method gives better classification results than hippocampal volume, but is outperformed by

Table 8.1: Demographic and clinical data of the study subjects. Level of significance is set to $p < 0.05$. *Different between the groups. [1]Different from controls. [2]Different from all other groups.

	N (F)	Age	MMSE*	Education	APOE ε4 carriers
CN	152 (35)	76.2±4.9	29.2±0.9[2]	16.1±2.7	29%
S-MCI	112 (28)	75.3±6.8	27.4±1.7[2]	16.1±3.0[1]	47%
P-MCI	110 (40)	74.6±6.8	26.7±1.7[2]	15.7±3.1[1]	67%
AD	103 (48)	75.5±7.2	23.2±2.0[2]	14.8±2.9[1]	69%

	ACC	SEN	SPE
HC	81	82	80
HA m12	79	81	76
HA m24	86	89	81
MBL bl	84	85	81
MBL m12	86	87	84
MBL m24	88	90	85
TBM bl	86	89	82
TBM m12	88	90	85
TBM m24	89	91	85
Combined bl	89	90	88
Combined m12	89	92	86
Combined m24	91	92	89

Table 8.2: CN vs AD. Accuracy (ACC), sensitivity (SEN) and specificity (SPE) are presented for hippocampal volume (HC), hippocampal atrophy (HA), manifold learning (MBL), tensor-based morphometry and the combination of all features. Features are available at baseline (BL), month 12 (m12) and month 24 (m24).

TBM. The combination of all the baseline measurements improved the results to 89% / 90% / 88% (ACC / SEN / SPE).

All individual features allowed improved classification accuracy when based on follow-up scans. It is remarkable that hippocampal atrophy, the only feature looking at intra-subject development, performs worse than hippocampal volume when measured over 12 months.

When using the combined features measured after / over 24 months, overall ACC / SEN / SPE improve to 91% / 92% / 89%.

S-MCI vs P-MCI classification

The classification results of S- and P-MCI subjects are presented in Table 8.3. With only small differences, the rank order of the baseline features from best to worst is manifold coordinates, hippocampal volume and TBM. The combination of baseline features improved the results by up to 8% units to 68% / 66% / 71% (ACC / SEN / SPE). Similar to CN vs AD classification, follow-up information improved classification rates to up to 72% / 70% / 74% with the combined use of features obtained from images acquired after 24 months.

	ACC	SEN	SPE
HC	66	65	66
HA m12	61	66	55
HA m24	65	65	65
MBL bl	66	67	65
MBL m12	67	66	69
MBL m24	68	68	68
TBM bl	63	65	61
TBM m12	66	68	65
TBM m24	69	69	68
Combined bl	68	66	71
Combined m12	71	70	74
Combined m24	72	70	74

Table 8.3: S-MCI vs P-MCI. Accuracy (ACC), sensitivity (SEN) and specificity (SPE) are presented for hippocampal volume (HC), hippocampal atrophy (HA), manifold learning (MBL), tensor-based morphometry and the combination of all features. Features are available at baseline (BL), month 12 (m12) and month 24 (m24).

CN vs P-MCI classification

Classification results of CN and P-MCI subjects are presented in Table 8.4. The rank order of the baseline features from best to worst is TBM, manifold coordinates and with some distance, hippocampal volume. Combination of baseline features improved results by up to 10% to 85% / 87% / 82% (ACC / SEN / SPE). The use of follow-up images improves results to up to 87% / 88% / 86% when using features after 24 months.

8.3.3 Classification results using dataset II

The restricted dataset with N=364 baseline images described above was used to apply all methods, including cortical thickness measurement. This dataset was used to perform classification on combinations of baseline features measured with different methods. Apart from the overall classification results, results for all possible combinations of features are presented. Such an analysis allows drawing conclusions on the influence of individual features on the classification accuracy obtained with a combined feature set.

Tables 8.5, 8.6 and 8.7 show the results for CN vs AD, S-MCI vs P-MCI and

	ACC	SEN	SPE
HC	77	79	76
HA m12	70	74	64
HA m24	77	81	72
MBL bl	81	83	79
MBL m12	82	84	80
MBL m24	83	86	79
TBM bl	83	87	77
TBM m12	84	86	81
TBM m24	86	87	85
Combined bl	85	87	82
Combined m12	86	87	85
Combined m24	87	88	86

Table 8.4: CN vs P-MCI. Accuracy (ACC), sensitivity (SEN) and specificity (SPE) are presented for hippocampal volume (HC), hippocampal atrophy (HA), manifold learning (MBL), tensor-based morphometry and the combination of all features. Features are available at baseline (BL), month 12 (m12) and month 24 (m24).

CN vs P-MCI respectively. When classifying stable from progressive MCI subjects, the overall best classification accuracy is achieved only when combining all available features. In both other comparisons, subsets of features are able to give equally good results to the whole set. In CN vs AD, TBM together with MBL or CTH performs as good as the whole feature set. In CN vs P-MCI, the combination of these three features achieves a performance as good as the full set of available features.

8.4 Discussion

In this chapter, the automatic diagnostic capabilities of 4 structural MRI features (CN, HA, MBL, TBM) was assessed separately and combined in a sample of 477 subjects with 2 years follow-up data from the ADNI database. In a restricted dataset with 364 subjects, cortical thickness was added as a fifth feature.

When applied separately to baseline features, TBM provided the overall best results for all the methods, closely followed by MBL. Combining all baseline methods improved the results in all study experiments. The use of follow-up images further enhanced classification accuracy by up to 6% in the S-MCI vs P-MCI classification. It shall, however, be noted that all features apart from hippocampal atrophy do not consider

	ACC	SEN	SPE
CTH	82	86	76
HC	80	79	82
MBL	86	88	83
TBM	88	89	87
MBL + HC	88	87	90
MBL + TBM	90	90	90
MBl + CTH	89	90	86
HC + TBM	88	88	87
HC + CTH	86	89	82
TBM + CTH	90	91	87
MBL + HC + TBM	90	90	90
MBL + HC + CTH	88	89	88
MBL + TBM + CTH	90	91	89
HC + TBM + CTH	89	90	88
All	90	90	89

Table 8.5: CN vs AD. Results for the combination of different feature sets are presented. The used feature sets include cortical thickness (CTH), hippocampal volume (HC), manifold learning (MBL) and tensor-based morphometry (TBM).

	ACC	SEN	SPE
CTH	63	65	60
HC	64	61	67
MBL	64	65	62
TBM	63	61	64
MBL + HC	64	65	61
MBL + TBM	63	61	65
MBl + CTH	64	66	62
HC + TBM	62	61	64
HC + CTH	65	67	64
TBM + CTH	63	65	61
MBL + HC + TBM	65	64	66
MBL + HC + CTH	64	65	62
MBL + TBM + CTH	65	64	66
HC + TBM + CTH	65	65	65
All	66	65	67

Table 8.6: S-MCI vs P-MCI. Results for the combination of different feature sets are presented. The used feature sets include cortical thickness (CTH), hippocampal volume (HC), manifold learning (MBL) and tensor-based morphometry (TBM).

	ACC	SEN	SPE
CTH	78	80	74
HC	77	76	78
MBL	81	85	74
TBM	81	84	75
MBL + HC	81	84	77
MBL + TBM	85	87	80
MBl + CTH	83	86	78
HC + TBM	81	83	79
HC + CTH	81	83	80
TBM + CTH	83	84	80
MBL + HC + TBM	83	87	78
MBL + HC + CTH	84	88	78
MBL + TBM + CTH	86	90	80
HC + TBM + CTH	83	84	80
All	86	89	80

Table 8.7: P-MCI vs CN. Results for the combination of different feature sets are presented. The used feature sets include cortical thickness (CTH), hippocampal volume (HC), manifold learning (MBL) and tensor-based morphometry (TBM).

actual intra-subject development. The reported improvements with follow-up data can therefore be mainly attributed to the pathomorphologically more advanced differences between the different subject groups. Such a development can be expected to be particularly significant for the S-MCI vs P-MCI comparison.

The presented results are in line with the results concerning single MRI methods in the CN/AD classification. Liu et al. reported SEN/SPE of 0.92/0.90 in the classification of CN/AD subjects using regional cortical volumes in the AddNeuroMed dataset [98]. In the presented study the results obtained with single methods are lower (0.80-0.89) but almost identical when the methods were combined. However, Liu and colleagues did not use cross-validation or separate training/testing sets when producing the results which could lead to overestimation of the results in a dataset outside the study cohort. Gerardin et al. [62] acquired a high SEN/SPE of 0.96/0.92 by using hippocampal shape analysis, but the number of subjects (25 CN, 23 AD) was quite low in order to produce results with good generalizability. Chupin et al. [30] reported a SEN / SPE of 75% / 77% (hippocampal volume) and Querbes et al. [114] an ACC of 85% (cortical thickness), both lower than the results acquired with the combination

of baseline features or TBM features independently in the presented study.

Varying results concerning AD prediction (S-MCI/P-MCI classification using the baseline measurements) have been published: Querbes et al. [114] reported an ACC of 73% (CTH analysis), Liu et al. [99] a SEN/SPE of 76%/68% (amygdala and caudate volumes), Chupin et al. [30] a SEN/SPE of 60%/65% (hippocampal volume) and Davatzikos et al. [39] a SEN/SPE of 95%/38% (SPAREAD index). The results with separate and combined baseline features presented here lie in the range of these results (SEN/SPE 67%/65%, 65%/66% and 66%/71% when using MBL, HC and the combined features, respectively).

There can be several explanations to the variation in the reported results. A majority of the studies in this field have used different statistical methods and MRI feature extraction strategies on different datasets, which makes a comparison of the results complicated. Also the variation in the size of the study samples and the use (or ignoring) of cross-validation or separate training/testing sets are important factors, which both have crucial impact on the reliability and generalizability of the results. Furthermore, since the ADNI study is still ongoing, several subjects labeled as S-MCI will progress in the future to the P-MCI group.

A recent study with a comparable dataset from ADNI assessed the classification performance of several structural MRI methods in experiments comparable to this study [38]. This is the biggest study of classification accuracy on ADNI with automated MR-based methods so far. A detailed comparison of the results reported in this study with the results obtained in this thesis is given in the conclusion, Chapter 9.

Some studies have also combined different biomarkers (CSF, MRI, PET) with the idea of measuring different aspects of AD pathology and thus improve the classification accuracy. Hinrichs et al. [74] improved their CN/AD classification ACC by a few % units to 81% by combining MRI and PET. Eckerström et al. [46] studied the separation of a unified CN/S-MCI group from P-MCI group with CSF proteins and manual hippocampal volumes. They found CSF to be superior to MRI (SEN/SPE 95%/79% vs 86%/66%) while the combination performed best (SEN/SPE 90%/91%).

However, it should be noted that the study sample in that particular study was small (a total of 68 subjects) and neither cross-validation or separate training/testing sets were used in order to ensure good generalizability of the results. In [89], the improvement from using multiple biomarkers was not significant and [39] reported marginal improvements which, however, may be related to the fact the results with only one biomarker where not very good already. In future work it might be interesting to see if different measurements are superior for different tasks, i.e., if a particular biomarker might be more useful for a certain subpopulation than another.

Considering the results of the presented study and those reported in literature, it seems questionable if the collection of several biomarkers or repeated examinations is worth the effort and resources. A combination of different features extracted from a single MRI seems to provide results that are comparable or better than those obtained with other or multiple biomarkers. However, the use of follow-up scans improves these numbers by only a fraction. In a clinical point of view, this is interesting since it means that a single MRI scan provides not only aid to differential diagnostics, but also reliably describes a persons phase in the CN/AD continuum. MRI is also widely available, non-invasive and often useful in the differential diagnostics of memory problems thus making it a compelling option as the first biomarker that would be obtained from a patient with mild memory problems.

Chapter 9

Summary and Conclusion

This thesis presents a detailed analysis on the extraction of biomarkers from brain MR images. From a methodological point of view, several novel approaches are described. From an application point of view, a rigorous evaluation on a large and diverse image dataset is presented.

The goal of this work was closely aligned to the project goals during which it was carried out. PredictAD aims at the development of a unified biomarker for AD that can be extracted in routine clinical use. Apart from accuracy, a main focus was therefore set on robustness with computational speed being another important factor. The ADNI study is the biggest study on MR imaging in dementia so far [111]. With its large number of participants, the use of dozens of different imaging sites and the use of equipment from all major scanner vendors, it provides a dataset that is close to what can be expected in clinical practice. The challenge of this work lied in the development of methods that can be robustly applied in a general fashion to such a dataset.

Chapters 3 and 4 presented a framework to apply multi-atlas segmentation in a robust and automated way to a diverse dataset. Previous work that combined multi-atlas segmentation with an intensity-based refinement step [140] was extended to be applicable in a fully-automated way. LEAP, a novel method to propagate a set of atlases in a step-wise fashion to a diverse dataset, was proposed, significantly improving

traditional multi-atlas segmentation. In Chapter 5, an extension of this method was described that provides a consistent segmentation of longitudinal datasets and allows an accurate measurement of atrophy.

Chapters 6 and 7 propose a new, data-driven approach for biomarker extraction. In the proposed framework, information derived from inter- and intra-subject imaging similarities can be combined with non-imaging metadata available for the study subjects. Both measures are combined to define a unified, low-dimensional manifold representation of the population. In this low-dimensional representation, neighborhoods represent similarity according to the measurements incorporated. Inferences can be made from subjects with known clinical status to subjects with no defined label.

Chapter 8 is motivated by the defined goal to obtain an accurate and reliable biomarker from combining different measurements. A comprehensive analysis of the biomarkers developed in this thesis in combination with other MR-based measures is presented.

9.1 Classification performance

With the main goal being the definition of biomarkers for AD, most of the evaluations presented in this work are based on the power of a particular biomarker to discriminate between clinically relevant subject groups. A comparison of different methods based on this aspect is generally difficult due to differences in datasets used for evaluation. With the ADNI database as a quasi standard in brain imaging for AD, however, a more objective comparison is possible: Table 9.1 shows sensitivity (SEN) and specificity (SPE) values for the classification between groups of interest when using the proposed methods. In addition, results from recent publications using established methods in neuroimaging are presented. Only studies that were applied to the complete ADNI dataset that was available at the time of publication are presented. Other studies that use artificially restricted subsets of ADNI are difficult to compare in an

objective way since inclusion / exclusion criteria are sometimes unclear. While the ADNI images available in the different studies are still different due to the difference in date between the studies, good overlap permits comparison between the methods based on classification performance.

Table 9.1: Comparison of classification results achieved with the proposed method to state-of-the art methods. I: AD vs CN, II: P-MCI vs S-MCI, III: P-MCI vs CN. SEN/SPE. Classification accuracy is reported where no SEN/SPE was provided.

Method	Ref.	Subjects	Feature (s)	I	II	III
LEAP: multi-atlas, int. refinement	Ch. 4 / [151]	796 BL	Volume for 83 structures	79/87	49/73	67/83
LEAP: multi-atlas, int. refinement	Ch. 4 / [151]	796 BL	Hippocampal volume	71/82	28/84	64/82
4D graph-cuts	Ch. 5 / [154]	362 BL, 362 M24	Hippocampal atrophy over 24 months	85/87	66/69	79/85
Manifold	Ch. 6 / [152]	420 BL	Manifold coordinates: image similarities	81/88	50/72	74/85
Manifold with metadata	Ch. 6 / [152]	420 BL	Manifold coordinates: image similarities and hippo. vol. / Aβ_{42}	85/90	65/70	82/88
Longitudinal manifold	Ch. 7 / [150]	362 BL, 362 M24	Manifold coordinates: intra- and inter-subject variation	82/88	64/70	76/88
Combination of multiple biomarkers	Ch. 8	477 BL	Hippo. vol., Manifold coord., tensor based morphometry	90/88	66/71	87/82
Voxel-based morphometry	[6, 7, 38]	509 BL	gray-matter (GM) tissue probabilities	81/95	0/100	57/96
STAND-score	[144, 38]	509 BL	Feature selection based on GM tissue probabilities	69/90	57/78	73/85
COMPARE	[51, 38]	509 BL	Feature selection based on GM tissue probabilities	66/86	62/67	49/81
Freesurfer	[52, 38]	509 BL	Cortical thickness	74/90	32/91	54/96
Freesurfer	[53, 38]	509 BL	Hippocampal vol.	63/80	61/70	73/74
Probabilistic atlas, hybrid constraints	[30]	605 BL	Hippocampal vol.	75/77	60/65	67/72
Multi-atlas, int. refinement	[103]	776 BL	Hippocampal vol.	80	63	n.a.
Spherical harmonics	[62, 38]	509 BL	Hippocampal shape	69/84	0/100	57/88

The discrimination between three pairings of subject groups is of main clinical interest: the discrimination of Alzheimer's subjects and subjects with progressive MCI (P-MCI) from healthy subjects, AD vs CN (listed I in Table 9.1) and P-MCI vs CN (II). The most challenging yet clinically most important discrimination is between progressive and stable (S-MCI) MCI subjects (III). A reliable detection of subjects at risk of converting from MCI allows to decide on potential disease modifying treatments.

The performance of the different methods in all three tasks is presented in Table 9.1. Based on only image similarities at baseline, the presented manifold framework together with voxel-based morphometry [6] and the STAND-score [144] outperform the methods based on traditional volumetry and morphometry. The biomarker extracted from manifold coordinates based on image similarities and metadata result in the best classification accuracy based on a single method. The overall best classification accuracy is obtained in the comprehensive analysis combining multiple different biomarkers described in Chapter 8 of this thesis.

The methods that use structural volumes for classification perform slightly worse than the ones based on machine learning. While no single method shows systematically the best results, the proposed LEAP method with 83 volumes gives the overall stablest results. Biomarkers incorporating longitudinal development show significantly better classification performance than the relevant baseline measure. The biggest improvement, however, is made in the detection of P-MCI subjects. Since these subjects undergo fast pathomorphological change, the improved classification accuracy may mainly be attributed to the more significant inter-subject differences after two years than the consideration of intra-subject development itself.

9.2 Performance based on other measures

For differential diagnosis, which is the main goal in PredictAD, classification performance of a given method is the most important metric to measure its performance. In other settings, other measures are of higher importance. In the following, several

such measures are discussed by comparing their performance on the proposed method to that of state-of-the-art methods.

9.2.1 Label overlaps

When purely evaluating novel methodology, the consistency of an automatically derived measure with some gold-standard is commonly used. For structural segmentation methods as the one presented in Chapter 4, the volume overlap between a manually extracted volume and some reference volume is commonly used as a performance measure. A widely used measure is the Dice overlap or Similarity Index (SI) [42]. It gives a value of 1 for a perfect overlap between the two volumes compared and a value of 0 for no overlap.

Table 9.2 compares the dice overlap for hippocampus segmentation achieved with the method presented in Chapter 4 to the results achieved with state-of-the-art automated approaches.

Table 9.2: Hippocampus label overlap

Method	Ref.	Subjects	SI
LEAP	Ch. 4 / [151]	182 (ADNI)	0.85 ± 0.03 (L) 0.85 ± 0.03 (R)
Multi-atlas	Heckemann [73]	30	0.81 ± 0.04 (L) 0.83 ± 0.04 (R)
Freesurfer: probabilistic atlas, intensity model	Fischl [53, 70]	13	0.87
AdaBoost based on gray image features	Morra [108]	21	0.86 (L) 0.85 (R)
Probabilistic atlas, hybrid constraints	Chupin [30]	16	0.87 ± 0.02
Multi-atlas, Int. model.	van der Lijn [140]	20	0.85 ± 0.04 (L) 0.86 ± 0.02 (R)
Multi-atlas with selection, Int. model.	Leung [96]	15 (ADNI)	0.93 (L)
Multi-atlas with selection, Int. model.	Lötjönen [103]	340 (ADNI)	0.87 ± 0.04

Apart from the relatively low SI value achieved with multi atlas-segmentation with no intensity refinement [73] and the results presented by Leung et al [96] with SI=0.93, all SI values presented in Table 9.2 lie in the range of 0.85-0.87. Since all evaluations

are performed on different datasets it is, however, difficult to directly compare the results. The manual delineation of a reference segmentation is time-intensive and expensive which restricts the set of reference labels in most studies to around 20-30. Often, these reference segmentations are based on only a subset of the whole brain population which makes statements about a method's robustness difficult. While the reference labels used in this thesis and by Lötjönen et al. [103] are based on a semi-automated protocol with manual correction, they cover a much wider population and are also available for more subjects. Nevertheless, the achieved label overlap with both methods is in the same area as the results based on a more restricted evaluation.

While the method proposed in [103] outperforms the presented LEAP method, it is evaluated under different conditions. In this method, atlas-selection is performed in multi-atlas label propagation where available atlases cover the whole spectrum of target images. The method presented in Chapter 4 on the other hand is a proof-of-principle for a framework that propagates a set of atlases based on a sub-population to a diverse image set. While such a framework can be applied to a diverse image set without the need of an atlas database that covers the whole population, it results in similar label overlaps as a method that specifically uses atlases tailored to the target population. The classification accuracy achieved with the LEAP method applied to 83 brain structures and presented in Chapter 4, outperforms the one achieved with the hippocampus segmentation of [103] (see Table 9.1). A conclusion from these observations is that the selection of suitable atlases from a large database gives the best results with multi-atlas segmentation. The proposed LEAP method, however, performs similarly well when only using a small atlas database and is able to produce significantly superior results in the unavailability of a large database.

9.2.2 Sample size

An important measure in clinical trials is the required sample size to measure a hypothetical treatment effect with a given method [47]. A benchmark defined in the ADNI study is the sample size needed in a two-arm study to detect a 25% change in annual

atrophy rate with a power of 80% and 5% significance [111]. In this setting, the sample size is defined by the relation between mean atrophy rate and its standard deviation.

Table 9.3 presents atrophy rates and resulting sample sizes for several recent studies on ADNI. The presented 4D-graph cuts method described in Chapter 5 is compared to a method were a semi-automatic segmentation method is applied to measuring hippocampal atrophy [124], the Boundary Shift Integral (BSI) [96] as well as Deformation Based Morphometry (DBM) [76].

Recent communication in the neuroimaging community [136, 76, 56] discusses the importance of relating atrophy rates in dementia to healthy atrophy. Fox et al. [56] stress that "sample sizes should ideally be calculated using the excess change over normal ageing" in order to avoid a potential bias of automated methods. Following this, sample sizes corrected for healthy ageing (SS II in Table 9.3) are presented alongside sample sizes resulting from non-corrected atrophy rates (SS I).

Table 9.3: Atrophy rates over 12 months and resulting sample sizes required to detect a 25% change in atrophy rate with 80% power and 5% significance. All results are based on ADNI data, sample sizes are reported for both arms (AD/MCI). Sample sizes are presented for measured atrophy rates (SS I) and atrophy rates corrected for healthy ageing (SS II). N: Number of subjects used.

Method, ROI	Ref.	N	Atrophy rate (CN/MCI/AD)	SS I	SS II
4D-graph cuts, hippocampus	Ch. 5 / [154]	555	0.85±1.59/2.34±2.12/3.85±1.99	67/206	110/508
SNT labels, hippocampus	Schuff [124]	449	0.87±5.63/2.6±4.51/4.4±5.88	448/1176	696/1705
BSI, hippocampus	Leung [57, 96]	682	1.05±1.81/2.77±2.53/4.63±2.78	90/209	151/542
DBM, temporal lobe	Hua [76]	431	0.6±0.8/0.8±0.8/1.2±0.8	112/251	446/4014

The presented sample sizes span a wide spectrum, where the proposed 4D graph-cuts method and the established boundary shift integral outperform the independent segmentation of several timepoints (SNT) and deformation based morphometry based on the temporal lobe. While the latter method achieves competitive results with the former methods when directly using measured atrophy rates, a clear difference can

be observed when normalizing for normal ageing. This highlights the importance of such a normalization step in order to avoid a bias in the estimation of sample sizes in clinical trials.

9.3 Conclusion

This thesis presented several novel methods for an automated extraction of biomarkers for Alzheimer's disease from brain magnetic resonance imaging. While the application and usability of the presented methods in a clinical environment needs yet to be shown, the rigorous evaluation on a large and diverse image set as the one presented in ADNI shows promising results towards this goal. All proposed methods are able to perform a state-of-the art automated analysis and enrich the neuroimaging community as shown by their publication in renowned journals and on international conferences.

The described multi-atlas segmentation method (Chapters 3 - 5) provides a unified framework to measure structural volume and atrophy in longitudinal MR sequences in a robust and generic way. The multi-atlas segmentation with automated intensity-refinement which forms the core of this method has been applied in a fast and robust way, with a run-time of 3-4 minutes on a standard desktop PC [104]. The step-wise segmentation with LEAP (Chapter 4) and the extension to 4D segmentation (Chapter 5) only require marginal computation time on top.

The novel manifold-based classification framework described in Chapters 6 and 7 forms a fast alternative to traditional biomarkers. The presented classification accuracy performs favorably to many well-established biomarkers. The computation including all pre-processing requires less than 10 minutes on a standard desktop PC.

The proven robustness of the described methods together with a short run-time makes them strong contenders for a potential application in a clinical setting.

9.4 Future work

In future, more work is planned on the presented manifold learning framework. Incorporating additional prior knowledge may further improve the achieved classification and regression results. When simultaneously embedding baseline and follow-up scans as described in Chapter 7.2.2 and illustrated in Figure 7.1 (a), information about the acquisition data of individual subjects could be used as prior knowledge when learning the low-dimensional space. In order to model the natural development of brain structures over time, a term that penalizes the abrupt deviation of a subject's movement from an initial trajectory can be introduced. This would avoid a "back-and-forth" movement of a subject that could be caused by errors in the pairwise similarity measure. Initial tests with an extended Sammon mapping objective [121] function that incorporates such constraints show promising results. Furthermore, combining the approaches described in Chapters 6 and 7 may allow to better model disease progression in a low-dimensional manifold. The incorporation of metadata (e.g. CSF measurement, hippocampal volume) at baseline *and* follow-up can be expected to further improve the quality of the learned biomarker. More generally, the use of other imaging modalities apart from structural MRI could be considered. Positron emission tomography (PET) imaging with the tracers FDG and especially PiB has been shown to provide powerful biomarkers for Alzheimer's disease [69]. Potentially less sensitive to the detection of the desease but less invasive and therefore of interest are diffusion tensor imaging (DTI) and functional MRI (fMRI) [69]. Combining the MRI-based features proposed in this thesis with measurements obtained from such imaging modalities can be expected to lead to a more poweful biomarker. Apart from the option to combine the obtained features before the classification step, the proposed manifold-learning framework (Chapters 6, 7) offers the possibility to combine the features before the manifold-learning step to obtain a unified biomarker.

Another area of future research is the definition of the similarity measure used to perform manifold learning. In most applications described in this thesis, image similarities where evaluated over anatomically defined regions of interest. A more data-

driven approach, where the region over which similarities is evaluated is determined according to the quality of the resulting manifold, is desirable. One possible approach could be to obtain a boosted similarity measure as done before for distance functions in k-nn classifiers [137, 3].

Chapter 10

Publications

10.1 Book Chapter

(1) P. Aljabar, R. Wolz, D. Rueckert. Manifold learning for medical image registration, segmentation and classification. *In* Machine Learning in Computer-Aided Diagnosis: Medical Imaging Intelligence and Analysis *IGI Global*, in press, 2011

10.2 Journal Publications

(2) R. Wolz, P. Aljabar, J. V. Hajnal, J. Lötjönen, D. Rueckert. Nonlinear Dimensionality Reduction Combining MR Imaging with Non-Imaging Information. *Revised to Medical Image Analysis*, 2011

(3) R. Wolz*, V. Julkunen*, J. Koikkalainen, E. Niskanen, D. Zhang, J. Mattila, D. Rueckert, H. Soininen, J. Lötjönen. Comprehensive Analysis of MRI Images in Early Diagnostics of Alzheimers Disease. *Submitted to Neuroimage*, 2011

(4) P. Aljabar, R. Wolz, L. Srinivasan, S. Counsell, M. Rutherford, A. Edwards, J. V. Hajnal, D. Rueckert. A combined manifold learning analysis of shape and appearance to characterize neonatal brain development. *Submitted to IEEE Transactions on Medical Imaging*, 2011

*Both authors contributed equally

(5) L. Risser, F.-X. Vialard, R. Wolz, M. Murgasova, D. D. Holm, D. Rueckert. Simultaneous Multiscale Registration using Large Deformation Diffeomorphic Metric Mapping. *IEEE Transactions on Medical Imaging*, in press, 2011

(6) J. Lötjönen, R. Wolz, J. Koikkalainen, V. Julkunen, L. Thurfjell, R. Lundqvist, G. Waldemar, H. Soininen, D. Rueckert. Fast and robust extraction of hippocampus from MR images for diagnostics of Alzheimers disease. *NeuroImage*, 56(1): 185-196, 2011

(7) R. Wolz, R.A. Heckemann, P. Aljabar, J.V. Hajnal, A. Hammers, J. Lötjönen and D. Rueckert. Measurement of hippocampal atrophy using 4D graph-cut segmentation: Application to ADNI. *NeuroImage*, 52(1):109-118, 2010

(8) J. Lötjönen, R. Wolz, J. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen, D. Rueckert. Fast and robust multi-atlas segmentation of brain magnetic resonance images. *NeuroImage*, 49(3):2352-2365, 2010

(9) R. Wolz, P. Aljabar, J.V. Hajnal, A. Hammers, D. Rueckert. LEAP: Learning Embeddings for Atlas Propagation. *NeuroImage*, 49(2):1316-1325, 2010

10.3 Conference Proceedings

(10) R. Guerrero, R. Wolz, D. Rueckert. Laplacian Eigenmaps Manifold Learning for Landmark Localization in Brain MR Images. *MICCAI 2011*, in press, Toronto, Canada, 2011

(11) R. Wolz, P. Aljabar, J. V. Hajnal, J. Lötjönen, Daniel Rueckert. Manifold-based classification incorporating subject metadata. *MIUA 2011*, in press, London, UK, 2011

(12) E. Janousova, M. Vounou, R. Wolz, K. Gray, D. Rueckert, G. Montana. Fast brain-wide search of highly discriminative regions in medical images: an application to Alzheimer's disease. *MIUA 2011*, in press, London, UK, 2011

(13) K. Gray, R. Wolz, S. Keihaninejad, R. Heckemann, P. Aljabar, A. Hammers, D. Rueckert. Regional Analysis of FDG-PET for the Classification of Alzheimer's Disease. *MIUA 2011*, in press, London, UK, 2011

(14) R. Wolz, P. Aljabar, J. V. Hajnal, J. Lötjönen, Daniel Rueckert. Manifold Learning Combining Imaging with Non-Imaging Information. *ISBI 2011*, Chicago, USA, In proc. p 960-963, 2011

(15) J. Lötjönen, R. Wolz, J. Koikkalainen, L. Thurfjell, R. Lundqvist, G. Waldemar, H. Soininen, D. Rueckert. Improved Generation of Probabilistic Atlases for the Expectation Maximization Classification. *ISBI 2011*, Chicago, USA, In proc. p 1839-1842, 2011

(16) K. Gray, R. Wolz, S. Keihaninejad, R. Heckemann, P. Aljabar, A. Hammers, D. Rueckert. Regional Analysis of FDG-PET for use in the Classification of Alzheimer's disease. *ISBI 2011*, Chicago, USA, In proc. p 100-103, 2011

(17) R. Wolz, P. Aljabar, J.V. Hajnal and D. Rueckert. Manifold learning for biomarker discovery in MR imaging. *Workshop on Machine Learning in Medical Imaging, MICCAI 2010*, LNCS 6357/2010:116-123, 2010

(18) P. Aljabar, R. Wolz, L. Srinivasan, S. Counsell, J.P. Boardman, M. Murgasova, V. Doria, M. Rutherford, A.D. Edwards, J.V. Hajnal and D. Rueckert. Combining morphological information in a manifold learning framework: Application to neonatal MRI. *MICCAI 2010*, LNCS 6362/2010:1-8, 2010

(19) L. Risser, F. Vialard, R. Wolz, D. Holm, D. Rueckert. Simultaneous fine and coarse diffeomorphic registration: Application to the atrophy measurement in alzheimer's disease. *MICCAI 2010*, LNCS 6362/2010:610-617, 2010

(20) R. Wolz, R.A. Heckemann, P. Aljabar, J.V. Hajnal, A. Hammers, J. Lötjönen and D. Rueckert. Measuring atrophy by simultaneous segmentation of serial MR images using 4-D graph-cuts. *ISBI 2010*, In proc. p 960-963, 2010

(21) R. Wolz, P. Aljabar, R. A. Heckemann, A. Hammers, D. Rueckert. Segmentation of Subcortical Structures and the Hippocampus in Brain MRI using Graph-Cuts and Subject-Specific A-Priori Information. *ISBI 2009*, In proc. p 470-473, 2009

10.4 Conference Abstracts

(22) R. Wolz, R.A. Heckemann, P. Aljabar, J.V. Hajnal, A. Hammers, J. Lötjönen and D. Rueckert. Using automatically determined atrophy rates to discrimate between clinical groups. *Conference on the Virtual Physiological Human 2010*, Brussels, 2010

(23) R. Wolz, R.A. Heckemann, P. Aljabar, J.V. Hajnal, A. Hammers, J. Lötjönen and D. Rueckert. Automatically determined hippocampal atrophy rates in ADNI: their usability to discriminate between clinical groups and to detect changes in atrophy rate. *Alzheimer's and Dementia*, Volume 6, Issue 4, Supplement 1, July 2010, Page S284

(24) L. Thurfjell, J. Lötjönen, R. Wolz, J. Koikkalainen, G. Waldemar, H. Soininen, D. Rueckert. Fast and robust segmentation of brain magnetic resonance images. *Alzheimer's and Dementia*, Volume 6, Issue 4, Supplement 1, July 2010, Page S34

(25) L. Thurfjell, R. Lundqvist, J. Lötjönen, J. Koikkalainen, R. Wolz, D. Rueckert, R. Vandenberghe, H. Soininen, G. Waldemar. Comparison of biomarkers from PET [18F]flutemetamol amyloid imaging and structural MRI. *Alzheimer's and Dementia*, Volume 6, Issue 4, Supplement 1, July 2010, Page S55

(26) R. Risinger, R. M. Berman, V. Coric, J. Han, S. Kaplita, L. Burns, Z. Bhagwagar, D. Hill, R. Wolz, D. Rueckert, H. Feldman. Mild to moderate Alzheimer's disease (AD) clinical trials: The role of hippocampal atrophy as an inclusion criterion. *Alzheimer's and Dementia*, Volume 6, Issue 4, Supplement 1, July 2010, Page S285-S286

(27) K. R. Gray, R. Wolz, R. A. Heckemann, A. Hammers, D. Rueckert. Classification of ADNI subjects based on longitudinal analysis of the hippocampal FDG-PET signal. *Alzheimer's and Dementia*, Volume 6, Issue 4, Supplement 1, July 2010, Page S288-S289

10.5 Patent

(28) LEAP: Method and apparatus for processing medical images. PCT/GB2010/001844, September 2010

Appendix A

ADNI

A.1 MR image acquisition

In the ADNI study, image acquisition was carried out at multiple sites based on a standardized MRI protocol [79] using 1.5T scanners manufactured by General Electric Healthcare (GE), Siemens Medical Solutions, and Philips Medical Systems. Out of two available 1.5T T1-weighted MR images based on a 3D MPRAGE sequence, we used the image that has been designated as "best" by the ADNI quality assurance team [79]. Acquisition parameters on the SIEMENS scanner (parameters for other manufacturers differ slightly) are echo time (TE) of 3.924 ms, repetition time (TR) of 8.916 ms, inversion time (TI) of 1000 ms, flip angle 8°, to obtain 166 slices of 1.2-mm thickness with a 256 × 256 matrix.

All images were preprocessed by the ADNI consortium using the following pipeline:

1. *GradWarp*: A system-specific correction of image geometry distortion due to gradient non-linearity [86].

2. *B1 non-uniformity correction*: Correction for image intensity non-uniformity [79].

3. *N3*: A histogram peak sharpening algorithm for bias field correction [126].

Since the Philips systems used in the study were equipped with B1 correction and

their gradient systems tend to be linear [79], the preprocessing steps 1. and 2. were applied by ADNI only to images acquired with GE and Siemens scanners.

A.1.1 Hippocampus reference labels

For a subset of images, ADNI provides reference hippocampus label maps. To define these label maps, semi-automated hippocampal volumetry was carried out using a commercially available high dimensional brain mapping tool (Medtronic Surgical Navigation Technologies, Louisville, CO), that has previously been compared to manual tracing of the hippocampus [75]. First, 22 control points were placed manually as local landmarks for the hippocampus on the individual brain MRI data: one landmark at the hippocampal head, one at the tail, and four per slice (i.e., at the superior, inferior, medial and lateral boundaries) on five equally spaced slices perpendicular to the long axis of the hippocampus. Second, fluid image transformation was used to match the individual brains to a template brain [26]. Transformed label maps were inspected and if necessary manually corrected by qualified reviewers. Empirically, we found that the resulting hippocampal delineations start anteriorly with their separation from the amygdalae; include the bulk of the hippocampal subfields CA1-4 [100], the subiculum, the dentate gyrus; miss some of the medial hippocampal head at the level of the uncus; but contain most of the intralimbic gyrus, the alveus as well as much of the fimbria, and end posteriorly shortly posterior to where cella media, temporal horn, and occipital horn fuse on coronal slices.

Appendix B

Hammers atlas

The 30 T1-weighted MR images used to define the Hammers atlas were acquired with a 1.5T GE MR-scanner using an inversion recovery prepared fast spoiled gradient recall sequence with the following parameters: TE/TR 4.2 ms (fat and water in phase)/15.5 ms, time of inversion (TI) 450 ms, flip angle 20°, to obtain 124 slices of 1.5-mm thickness with a field of view of 18 × 24 cm with a 192 × 256 image matrix.

Table B.1 gives an overview on the 83 structures defined in each atlas image.

Structure	No right	No left
Hippocampus	1	2
Amygdala	3	4
Anterior temporal lobe, medial part	5	6
Anterior temporal lobe, lateral part	7	8
Gyri parahippocampalis et ambiens	9	10
Superior temporal gyrus, posterior part	11	12
Medial and inferior temporal gyri	13	14
Lateral occipitotemporal gyrus, gyrus fusiformis	15	16
Cerebellum	17	18
Brainstem, spans the midline	19	
Insula	21	20
Occipital lobe	23	22
Cingulate gyrus, anterior part	25	24
Cingulate gyrus, posterior part	27	26
Frontal lobe left, becomes middle frontal gyrus after subdivision of frontal	29	28
Posterior temporal lobe	31	30
Parietal lobe	33	32
Caudate nucleus	35	34
Nucleus accumbens	37	36
Putamen	39	38
Thalamus	41	40
Pallidum, globus pallidus	43	42
Corpus callosum	44	
Lateral ventricle, frontal horn, central part and occipital horn	45	46
Lateral ventricle, temporal horn	47	48
Third ventricle	49	
Precentral gyrus	51	50
Straight gyrus, gyrus rectus	53	52
Anterior orbital gyrus	55	54
Inferior frontal gyrus	57	56
Superior frontal gyrus	59	58
Postcentral gyrus	61	60
Superior parietal gyrus	63	62
Lingual gyrus	65	64
Cuneus	67	66
Medial orbital gyrus	69	68
Lateral orbital gyrus	71	70
Posterior orbital gyrus	73	72
Substantia nigra	75	74
Subgenual frontal cortex	77	76
Subcallosal area	79	78
Pre-subgenual frontal cortex	81	80
Superior temporal gyrus, anterior part	83	82

Table B.1: 83 Structures

Bibliography

[1] P. Aljabar, R. Heckemann, A. Hammers, J. Hajnal, and D. Rueckert. Multi-atlas based segmentation of brain images: Atlas selection and its effect on accuracy. *NeuroImage*, 46(3):726 – 738, 2009.

[2] P. Aljabar, D. Rueckert, and W. Crum. Automated morphological analysis of magnetic resonance brain imaging using spectral analysis. *NeuroImage*, 43(2):225 – 235, 2008.

[3] J. Amores, N. Sebe, and P. Radeva. Boosting the distance estimation: Application to the k-nearest neighbor classifier. *Pattern Recognition Letters*, 27(3):201 – 209, 2006.

[4] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de Solorzano. Combination Strategies in Multi-Atlas Image Segmentation: Application to Brain MR Data. *IEEE Transactions on Medical Imaging*, 28(8):1266 –1277, 2009.

[5] J. Ashburner and K. J. Friston. Nonlinear spatial normalization using basis functions. *Human Brain Mapping*, 7(4):254–266, 1999.

[6] J. Ashburner and K. J. Friston. Voxel-based morphometry - the methods. *NeuroImage*, 11(6):805–821, 2000.

[7] J. Ashburner and K. J. Friston. Unified segmentation. *NeuroImage*, 26(3):839–851, 2005.

[8] J. Ashburner, C. Hutton, R. Frackowiak, I. Johnsrude, C. Price, and K. Friston. Identifying global anatomical differences: Deformation-based morphometry. *Human Brain Mapping*, 6(5-6):348–357, 1998.

[9] B. Avants and J. C. Gee. Geodesic estimation for large deformation anatomical shape averaging and interpolation. *NeuroImage*, 23(Supplement 1):S139 – S150, 2004.

[10] K. O. Babalola, T. F. Cootes, and C. J. T. et al. 3D brain segmentation using active appearance models and local regressors. In *MICCAI 2008*, pages 401–408, 2008.

[11] R. Bajcsy, R. Lieberson, and M. Reivich. A computerized system for the elastic matching of deformed radiographic images to idealized atlas images. *Journal of Computer Assisted Tomography*, 7:618–625, Aug. 1983.

[12] J. Barnes, J. W. Bartlett, L. A. van de Pol, C. T. Loy, R. I. Scahill, C. Frost, P. Thompson, and N. C. Fox. A meta-analysis of hippocampal atrophy rates in alzheimer's disease. *Neurobiology of Aging*, 30(11):1711 – 1723, 2009.

[13] J. Barnes, J. Foster, R. Boyes, T. Pepple, E. Moore, J. Schott, C. Frost, R. Scahill, and N. Fox. A comparison of methods for the automated calculation of volumes and atrophy rates in the hippocampus. *NeuroImage*, 40(4):1655 – 1671, 2008.

[14] M. Belkin and P. Niyogi. Laplacian eigenmaps for dimensionality reduction and data representation. *Neural Computation*, 15(6):1373–1396, 2003.

[15] K. K. Bhatia, P. Aljabar, J. P. Boardman, L. Srinivasan, M. Murgasova, S. J. Counsell, M. A. Rutherford, J. V. Hajnal, A. D. Edwards, and D. Rueckert. Groupwise combined segmentation and registration for atlas construction. In *MICCAI (1)*, volume 4791 of *Lecture Notes in Computer Science*, pages 532–540. Springer, 2007.

[16] D. J. Blezek and J. V. Miller. Atlas stratification. *Medical Image Analysis*, 11(5):443 – 457, 2007.

[17] R. G. Boyes, D. Rueckert, P. Aljabar, J. Whitwell, J. M. Schott, D. L. Hill, and N. C. Fox. Cerebral atrophy measurements using Jacobian integration: Comparison with the boundary shift integral. *NeuroImage*, 32(1):159 – 169, 2006.

[18] Y. Boykov, O. Veksler, and R. Zabih. Fast approximate energy minimization via graph cuts. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(11):1222–1239, Nov 2001.

[19] H. Braak and E. Braak. Neuropathological stageing of Alzheimer-related changes. *Acta Neuropathologica*, 82(4):239 – 259, 1991.

[20] R. Brookmeyer, E. Johnson, K. Ziegler-Graham, and H. M. Arrighi. Forecasting the global burden of Alzheimer's disease. *Alzheimer's and Dementia*, 3(3):186 – 191, 2007.

[21] C. C. Brun, N. Lepor, X. Pennec, A. D. Lee, M. Barysheva, S. K. Madsen, C. Avedissian, Y.-Y. Chou, G. I. de Zubicaray, K. L. McMahon, M. J. Wright, A. W. Toga, and P. M. Thompson. Mapping the regional influence of genetics on brain structure variability – a tensor-based morphometry study. *NeuroImage*, 48(1):37 – 49, 2009.

[22] C. J. C. Burges. A tutorial on support vector machines for pattern recognition. In *Knowledge Discovery and Data Mining*, volume 2, pages 121–167. 1998.

[23] W.-Y. Chang, C.-S. Chen, and Y.-P. Hung. Analyzing facial expression by fusing manifolds. In *Lecture Notes in Computer Vision - ACCV 2007*, pages 621–630. 2007.

[24] H. T. Chen, H. W. Chang, and T. L. Liu. Local discriminant embedding and its variants. In *Computer Vision and Pattern Recognition*, pages II: 846–853, 2005.

[25] G. Chetelat, B. Desgranges, V. De La Sayette, F. Viader, F. Eustache, and J. Baron. Mild cognitive impairment: Can FDG-PET predict who is to rapidly convert to Alzheimer's disease? *Neurology*, 60(8):1374–7, 2003.

[26] G. E. Christensen, S. C. Joshi, and M. I. Miller. Volumetric transformation of brain anatomy. *IEEE Transactions on Medical Imaging*, 16(6):864–877, 1997.

[27] G. E. Christensen, R. D. Rabbitt, and M. I. Miller. 3D brain mapping using a deformable neuroanatomy. *Physics in Medicine and Biology*, 39(3):609, 1994.

[28] F. R. K. Chung. Spectral graph theory. *Regional Conference Series in Mathematics, American Mathematical Society*, 92:1–212, 1997.

[29] M. Chupin, E. Gerardin, R. Cuingnet, C. Boutet, L. Lemieux, S. Lehericy, H. Benali, L. Garnero, and O. Colliot. Fully automatic hippocampus segmentation and classification in Alzheimer's disease and mild cognitive impairment applied on data from ADNI. *Hippocampus*, 19(6):579 – 587, 2009.

[30] M. Chupin, A. Hammers, R. Liu, O. Colliot, J. Burdett, E. Bardinet, J. Duncan, L. Garnero, and L. Lemieux. Automatic segmentation of the hippocampus and the amygdala driven by hybrid constraints: Method and validation. *NeuroImage*, 46(3):749 – 761, 2009.

[31] M. Chupin, A. R. Mukuna-Bantumbakulu, D. Hasboun, E. Bardinet, S. Baillet, S. Kinkingnhun, L. Lemieux, B. Dubois, and L. Garnero. Anatomically constrained region deformation for the automated segmentation of the hippocampus and the amygdala: Method and validation on controls and patients with alzheimer's disease. *NeuroImage*, 34(3):996 – 1019, 2007.

[32] D. L. Collins, C. J. Holmes, T. M. Peters, and A. C. Evans. Automatic 3-D model-based neuroanatomical segmentation. *Human Brain Mapping*, 3(3):190–208, 1995.

[33] D. L. Collins, A. P. Zijdenbos, W. F. C. Baar, and A. C. Evans. ANIMAL+INSECT: Improved cortical structure segmentation. In *Information Processing in Medical Imaging*, pages 210–223. Springer, 1999.

[34] J. Costa and A. Hero III. Classification constrained dimensionality reduction. *International Conference on Acoustics, Speech and Signal Processing*, pages 1077 – 1080, 2005.

[35] T. Cover and P. Hart. Nearest neighbor pattern classification. *IEEE Transactions on Information Theory*, 13(1):21–27, 1967.

[36] T. F. Cox and M. A. A. Cox. *Multidimensional Scaling*. Chapman & Hall, London, 1994.

[37] W. R. Crum, P. A. Freeborough, and N. C. Fox. The use of regional fast fluid registration of serial MRI to quantify local change in neurodegenerative disease. In *Medical Image Understanding and Analysis 1999*, 1999.

[38] R. Cuingnet, E. Gerardin, J. Tessieras, G. Auzias, S. Lehricy, M.-O. Habert, M. Chupin, H. Benali, and O. Colliot. Automatic classification of patients with Alzheimer's disease from structural MRI: A comparison of ten methods using the ADNI database. *NeuroImage*, In Press, Corrected Proof:–, 2010.

[39] C. Davatzikos, P. Bhatt, L. M. Shaw, K. N. Batmanghelich, and J. Q. Trojanowski. Prediction of MCI to AD conversion, via MRI, CSF biomarkers, and pattern classification. *Neurobiology of Aging*, In Press, Corrected Proof:–, 2010.

[40] B. C. Davis, P. T. Fletcher, E. Bullitt, and S. Joshi. Population shape regression from random design data. In *International Conference on Computer Vision*, pages 1–7, 2007.

[41] D. P. Devanand, G. Pradhaban, X. Liu, A. Khandji, S. De Santi, S. Segal, H. Rusinek, G. H. Pelton, L. S. Honig, R. Mayeux, Y. Stern, M. H. Tabert,

and M. J. de Leon. Hippocampal and entorhinal atrophy in mild cognitive impairment - Prediction of Alzheimer disease. *Neurology*, 68(11):828–836+, 2007.

[42] L. R. Dice. Measures of the amount of ecologic association between species. *Ecology*, 26(3):297–302, 1945.

[43] D. L. Donoho and C. Grimes. Hessian eigenmaps: New locally linear embedding techniques for high-dimensional data. Technical report, 2003.

[44] N. R. Draper and H. Smith. *Applied Regression Analysis (Wiley Series in Probability and Statistics)*. Wiley, 1998.

[45] B. Dubois, H. H. Feldman, C. Jacova, S. T. DeKosky, P. Barberger-Gateau, J. Cummings, A. Delacourte, D. Galasko, S. Gauthier, G. Jicha, K. Meguro, J. O'Brien, F. Pasquier, P. Robert, M. Rossor, S. Salloway, Y. Stern, P. J. Visser, and P. Scheltens. Research criteria for the diagnosis of Alzheimer's disease: revising the NINCDS-ADRDA criteria. *The Lancet Neurology*, 6(8):734 – 746, 2007.

[46] C. Eckerstrom, U. Andreasson, E. Olsson, S. Rolstad, K. Blennow, H. Zetterberg, H. Malmgren, A. Edman, and A. Wallin. Combination of hippocampal volume and cerebrospinal fluid biomarkers improves predictive value in mild cognitive impairment. *Dement Geriatr Cogn Disord*, 29(4):294–300, 2010.

[47] J. Eng. Sample size estimation: how many individuals should be studied? *Radiology*, 227(2):309–313, 2003.

[48] A. Ericsson, P. Aljabar, and D. Rueckert. Construction of a patient-specific atlas of the brain: Application to normal aging. In *IEEE International Symposium on Biomedical Imaging*, pages 480–483, 2008.

[49] A. C. Evans, J. P. Lerch, J. Pruessner, A. P. Zijdenbos, S. J. Teipel, and H. Hampel. Cortical thickness in Alzheimer's disease. *Alzheimer's and Dementia*, 1(1, Supplement 1):7 – 7, 2005.

[50] Y. Fan, N. Batmanghelich, C. M. Clark, and C. Davatzikos. Spatial patterns of brain atrophy in MCI patients, identified via high-dimensional pattern classification, predict subsequent cognitive decline. *NeuroImage*, 39(4):1731 – 1743, 2008.

[51] Y. Fan, D. Shen, R. C. Gur, R. E. Gur, and C. Davatzikos. COMPARE: Classification of morphological patterns using adaptive regional elements. *IEEE Transactions on Medical Imaging*, 26(1):93–105, 2007.

[52] B. Fischl and A. Dale. Measuring the thickness of the human cerebral cortex from magnetic resonance images. *Proceedings of the National Academy of Sciences*, 97:11044–11049, 2000.

[53] B. Fischl, D. H. Salat, E. Busa, M. Albert, M. Dieterich, C. Haselgrove, A. van der Kouwe, R. Killiany, D. Kennedy, S. Klaveness, A. Montillo, N. Makris, B. Rosen, and A. M. Dale. Whole brain segmentation: Automated labeling of neuroanatomical structures in the human brain. *Neuron*, 33(3):341 – 355, 2002.

[54] M. F. Folstein, S. E. Folstein, and P. R. McHugh. Mini-mental state: A practical method for grading the cognitive state of patients for the clinician. *Journal of Psychiatric Research*, 12(3):189–198, 1975.

[55] N. C. Fox, S. Cousens, R. Scahill, R. J. Harvey, and M. N. Rossor. Using serial registered brain magnetic resonance imaging to measure disease progression in Alzheimer disease: Power calculations and estimates of sample size to detect treatment effects. *Arch Neurol*, 57(3):339–344, 2000.

[56] N. C. Fox, G. R. Ridgway, and J. M. Schott. Algorithms, atrophy and alzheimer's disease: Cautionary tales for clinical trials. *NeuroImage*, In Press, Accepted Manuscript:–, 2011.

[57] P. A. Freeborough and N. C. Fox. The boundary shift integral: An accurate and robust measure of cerebral volume changes from registered repeat MRI. *IEEE Transactions on Medical Imaging*, 16(5):623–629, Oct 1997.

[58] P. A. Freeborough and N. C. Fox. Modeling brain deformations in alzheimer disease by fluid registration of serial 3D MR images. *Journal of Computer Assisted Tomography*, 22(5):838–43, 1998.

[59] Y. Freund and R. Schapire. A short introduction to boosting. *Journal of Japanese Society for Artificial Intelligence*, 14(5):771–780, 1999.

[60] J. C. Gee, M. Reivich, and R. Bajcsy. Elastically deforming 3D atlas to match anatomical brain images. *Journal of Computer Assisted Tomography*, 17(2):225–236, 1993.

[61] C. R. Genovese, N. A. Lazar, and T. Nichols. Thresholding of statistical maps in functional neuroimaging using the false discovery rate. *NeuroImage*, 15(4):870–878, 2002.

[62] E. Gerardin, G. Chetelat, M. Chupin, R. Cuingnet, B. Desgranges, H.-S. Kim, M. Niethammer, B. Dubois, S. Lehericy, L. Garnero, F. Eustache, and O. Colliot. Multidimensional classification of hippocampal shape features discriminates Alzheimer's disease and mild cognitive impairment from normal aging. *NeuroImage*, 47(4):1476 – 1486, 2009.

[63] S. Gerber, T. Tasdizen, P. T. Fletcher, S. Joshi, and R. Whitaker. Manifold modeling for brain population analysis. *Medical Image Analysis*, 14(5):643 – 653, 2010.

[64] I. S. Gousias, D. Rueckert, R. A. Heckemann, L. E. Dyet, J. P. Boardman, A. D. Edwards, and A. Hammers. Automatic segmentation of brain mris of 2-year-olds into 83 regions of interest. *NeuroImage*, 40(2):672 – 684, 2008.

[65] D. M. Greig, B. T. Porteous, and A. H. Seheult. Exact maximum a-posteriori estimation for binary images. *Journal of the Royal Statistical Society Series B-Methodological*, 51(2):271–279, 1989.

[66] J. Hamm, D. H. Ye, R. Verma, and C. Davatzikos. GRAM: A framework for geodesic registration on anatomical manifolds. *Medical Image Analysis*, 14(5):633 – 642, 2010.

[67] A. Hammers, R. Allom, M. J. Koepp, S. L. Free, R. Myers, L. Lemieux, T. N. Mitchell, D. J. Brooks, and J. S. Duncan. Three-dimensional maximum probability atlas of the human brain, with particular reference to the temporal lobe. *Human Brain Mapping*, 19(4):224–247, 2003.

[68] A. Hammers, R. Heckemann, M. J. Koepp, J. S. Duncan, J. V. Hajnal, D. Rueckert, and P. Aljabar. Automatic detection and quantification of hippocampal atrophy on MRI in temporal lobe epilepsy: A proof-of-principle study. *NeuroImage*, 36(1):38 – 47, 2007.

[69] H. Hampel, K. Brger, S. J. Teipel, A. L. Bokde, H. Zetterberg, and K. Blennow. Core candidate neurochemical and imaging biomarkers of alzheimer's disease. *Alzheimer's and Dementia*, 4(1):38 – 48, 2008.

[70] X. Han and B. Fischl. Atlas renormalization for improved brain MR image segmentation across scanner platforms. *IEEE Transactions on Medical Imaging*, 26(4):479–486, Apr. 2007.

[71] D. Harold, R. Abraham, P. Hollingworth, R. Sims, A. Gerrish, M. L. Hamshere, J. S. S. Pahwa, V. Moskvina, and K. Dowzell et. al. Genome-wide association study identifies variants at CLU and PICALM associated with Alzheimer's disease. *Nature Genetics*, 41:1088 – 1093, 2009.

[72] X. He, S. Yan, Y. Hu, P. Niyogi, and H.-J. Zhang. Face recognition using laplacianfaces. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 27:328–340, 2005.

[73] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers. Automatic anatomical brain MRI segmentation combining label propagation and decision fusion. *NeuroImage*, 33(1):115 – 126, 2006.

[74] C. Hinrichs, V. Singh, G. Xu, and S. Johnson. Mkl for robust multi-modality ad classification. In *MICCAI (II)*, volume 5762 of *Lecture Notes in Computer Science*, pages 786–794. Springer, 2009.

[75] Y.-Y. Hsu, N. Schuff, A.-T. Du, K. Mark, X. Zhu, D. Hardin, and M. W. Weiner. Comparison of automated and manual MRI volumetry of hippocampus in normal aging and dementia. *Journal of Magnetic Resonance Imaging*, 16(3):305–310, 2002.

[76] X. Hua, B. Gutman, C. Boyle, P. Rajagopalan, A. D. Leow, I. Yanovsky, A. R. Kumar, A. W. Toga, C. R. J. Jr., N. Schuff, G. E. Alexander, K. Chen, E. M. Reiman, M. W. Weiner, and P. M. Thompson. Accurate measurement of brain changes in longitudinal MRI scans using tensor-based morphometry. *NeuroImage*, In Press, Accepted Manuscript:–, 2011.

[77] X. Hua, A. D. Leow, N. Parikshak, S. Lee, M.-C. Chiang, A. W. Toga, C. R. J. Jr, M. W. Weiner, and P. M. Thompson. Tensor-based morphometry as a neuroimaging biomarker for alzheimer's disease: An mri study of 676 ad, mci, and normal subjects. *NeuroImage*, 43(3):458 – 469, 2008.

[78] M. D. Ikonomovic, W. E. Klunk, E. E. Abrahamson, C. A. Mathis, J. C. Price, N. D. Tsopelas, B. J. Lopresti, S. Ziolko, W. Bi, W. R. Paljug, M. L. Debnath, C. E. Hope, B. A. Isanski, R. L. Hamilton, and S. T. DeKosky. Post-mortem correlates of in vivo PiB-PET amyloid imaging in a typical case of Alzheimer's disease. *Brain*, 131(6):1630–1645.

[79] C. R. Jack Jr., M. A. Bernstein, N. C. Fox, P. Thompson, G. Alexander, D. Harvey, B. Borowski, P. J. Britson, J. L. Whitwell, C. Ward, A. M. Dale, J. P. Felmlee, J. L. Gunter, D. L. Hill, R. Killiany, N. Schuff, S. Fox-Bosetti, C. Lin, C. Studholme, C. S. DeCarli, G. Krueger, H. A. Ward, G. J. Metzger, K. T. Scott, R. Mallozzi, D. Blezek, J. Levy, J. P. Debbins, A. S. Fleisher, M. Albert, R. Green, G. Bartzokis, G. Glover, J. Mugler, and M. W. Weiner. The

Alzheimer's disease neuroimaging initiative (ADNI): MRI methods. *Journal of Magnetic Resonance Imaging*, 27(4):685–691, 2008.

[80] C. R. Jack Jr., R. C. Petersen, Y. C. Xu, P. C. O'Brien, G. E. Smith, R. J. Ivnik, B. F. Boeve, S. C. Waring, E. G. Tangalos, and E. Kokmen. Prediction of AD with MRI-based hippocampal volume in mild cognitive impairment. *Neurology*, 52(7):1397–1407, 1999.

[81] C. R. Jack Jr., R. C. Petersen, Y. C. Xu, S. C. Waring, P. C. O'Brien, E. G. Tangalos, G. E. Smith, R. J. Ivnik, and E. Kokmen. Medial temporal atrophy on MRI in normal aging and very mild Alzheimer's disease. *Neurology*, 49(3):786–794, 1997.

[82] C. R. Jack Jr., M. M. Shiung, J. L. Gunter, P. C. O'Brien, S. D. Weigand, D. S. Knopman, B. F. Boeve, R. J. Ivnik, G. E. Smith, R. H. Cha, E. G. Tangalos, and R. C. Petersen. Comparison of different MRI brain atrophy rate measures with clinical disease progression in AD. *Neurology*, 62(4):591–600, 2004.

[83] H. Jia, G. Wu, Q. Wang, and D. Shen. ABSORB: Atlas building by self-organized registration and bundling. *NeuroImage*, 51(3):1057 – 1070, 2010.

[84] I. T. Jolliffe. *Principal Component Analysis*. Springer-Verlag, 1986.

[85] S. Joshi, B. Davis, M. Jomier, and G. Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23(Supplement 1):151 – 160, 2004.

[86] J. Jovicich, S. Czanner, D. Greve, E. Haley, A. van der Kouwe, R. Gollub, D. Kennedy, F. Schmitt, G. Brown, J. MacFall, B. Fischl, and A. Dale. Reliability in multi-site structural MRI studies: Effects of gradient non-linearity correction on phantom and human data. *NeuroImage*, 30(2):436 – 443, 2006.

[87] V. Julkunen, E. Niskanen, S. Muehlboeck, M. Pihlajamki, M. Knnen, M. Hallikainen, M. Kivipelto, S. Tervo, R. Vanninen, A. Evans, and H. Soininen.

Cortical thickness analysis to detect progressive mild cognitive impairment: a reference to Alzheimer's disease. *Dementia and Geriatric Cognitive Disorders*, 28(5):404–412, 2009.

[88] S. Klein, M. Loog, F. van der Lijn, T. den Heijer, A. Hammers, M. de Bruijne, A. van der Lugt, R. Duin, M. M. B. Breteler, and W. Niessen. Early diagnosis of dementia based on intersubject whole-brain dissimilarities. In *IEEE International Symposium on Biomedical Imaging*, pages 249–252, 2010.

[89] O. Kohannim, X. Hua, D. P. Hibar, S. Lee, Y.-Y. Chou, A. W. Toga, C. R. J. Jr., M. W. Weiner, and P. M. Thompson. Boosting power for clinical trials using classifiers based on multiple biomarkers. *Neurobiology of Aging*, 31(8):1429 – 1442, 2010.

[90] W. J. Krzanowski. *Principles of Multivariate Analysis: A User's Perspective*. Oxford University Press, 1988.

[91] S. Landau, D. Harvey, C. Madison, E. Reiman, N. Foster, P. Aisen, R. Petersen, L. Shaw, J. Trojanowski, C. Jack Jr, M. Weiner, and W. Jagust. Comparing predictors of conversion and decline in mild cognitive impairment. *Neurology*, 75(3):230–8, 2010.

[92] M. Lehtovirta, H. Soininen, M. P. Laakso, K. Partanen, S. Helisalmi, A. Mannermaa, M. Ryynnen, P. H. J Kuikka, and S. P J Riekkinen. SPECT and MRI analysis in Alzheimer's disease: relation to apolipoprotein E epsilon 4 allele. *Journal of Neurology, Neurosurgery, and Psychiatry*, 60:644 – 649, 1996.

[93] A. D. Leow, I. Yanovsky, M. C. Chiang, A. D. Lee, A. D. Klunder, A. Lu, J. T. Becker, S. W. Davis, A. W. Toga, and P. M. Thompson. Statistical properties of jacobian maps and the realization of unbiased large-deformation nonlinear image registration. *IEEE Transactions on Medical Imaging*, 26(6):822–832, 2007.

[94] N. Lepore, C. Brun, Y. yu Chou, A. D. Lee, G. I. D. Zubicaray, M. Meredith, K. L. Mcmahon, M. J. Wright, A. W. Toga, and P. M. Thompson. Multi-atlas

tensor-based morphometry and its application to a genetic study of 92 twins. In *International Workshop on the Mathematical Foundations of Computational Anatomy (MFCA-2008)*, 2009.

[95] J. Lerch and A. Evans. Cortical thickness analysis examined through power analysis and a population simulation. *NeuroImage*, 24(1):163–173, 2005.

[96] K. K. Leung, J. Barnes, G. R. Ridgway, J. W. Bartlett, M. J. Clarkson, K. Macdonald, N. Schuff, N. C. Fox, and S. Ourselin. Automated cross-sectional and longitudinal hippocampal volume measurement in mild cognitive impairment and alzheimer's disease. *NeuroImage*, 51(4):1345 – 1359, 2010.

[97] K. K. Leung, M. J. Clarkson, J. W. Bartlett, S. Clegg, C. R. Jack Jr., M. W. Weiner, N. C. Fox, and S. Ourselin. Robust atrophy rate measurement in Alzheimer's disease using multi-site serial MRI: Tissue-specific intensity normalization and parameter selection. *NeuroImage*, 50(2):516 – 523, 2010.

[98] Y. Liu, T. Paajanen, Y. Zhang, E. Westman, L.-O. Wahlund, A. Simmons, C. Tunnard, T. Sobow, P. Mecocci, M. Tsolaki, B. Vellas, S. Muehlboeck, A. Evans, C. Spenger, S. Lovestone, and H. Soininen. Combination analysis of neuropsychological tests and structural MRI measures in differentiating AD, MCI and control groups–The AddNeuroMed study. *Neurobiology of Aging*, In Press, Corrected Proof:–, 2009.

[99] Y. Liu, T. Paajanen, Y. Zhang, E. Westman, L.-O. Wahlund, A. Simmons, C. Tunnard, T. Sobow, P. Mecocci, M. Tsolaki, B. Vellas, S. Muehlboeck, A. Evans, C. Spenger, S. Lovestone, and H. Soininen. Analysis of regional MRI volumes and thicknesses as predictors of conversion from mild cognitive impairment to Alzheimer's disease. *Neurobiology of Aging*, 31(8):1375 – 1385, 2010.

[100] R. Lorente de No. Studies on the Cerebral Cortex. II. Continuation of the Study of the Ammonic System. *Journal Psychology Neurology*, 46:113 – 190, 1934.

[101] P. Lorenzen, M. Prastawa, B. Davis, G. Gerig, E. Bullitt, and S. Joshi. Multi-modal image set registration and atlas formation. *Medical Image Analysis*, 10(3):440 – 451, 2006. Special Issue on The Second International Workshop on Biomedical Image Registration (WBIR'03).

[102] J. Lotjonen, J. Koikkalainen, L. Thurfjell, and D. Rueckert. Atlas-based registration parameters in segmenting sub-cortical regions from brain mri-images. In *Biomedical Imaging: From Nano to Macro, 2009. ISBI '09. IEEE International Symposium on*, 28 2009.

[103] J. Lotjonen, R. Wolz, J. Koikkalainen, V. Julkunen, L. Thurfjell, R. Lundqvist, G. Waldemar, H. Soininen, and D. Rueckert. Fast and robust extraction of hippocampus from mr images for diagnostics of alzheimer's disease. *NeuroImage*, In Press, Accepted Manuscript:–, 2011.

[104] J. M. Lotjonen, R. Wolz, J. R. Koikkalainen, L. Thurfjell, G. Waldemar, H. Soininen, and D. Rueckert. Fast and robust multi-atlas segmentation of brain magnetic resonance images. *NeuroImage*, 49(3):2352 – 2365, 2010.

[105] J. Mattila, J. Koikkalainen, M. van Gils, J. Lotjonen, G. Waldemar, A. Simonsen, D. Rueckert, L. Thurfjell, and H. Soininen. PredictAD - a clinical decision support system for early diagnosis of Alzheimer's disease. In *1st Virtual Physiological Human Conference*, pages 148–150, 2010.

[106] J. C. Mazziotta, A. W. Toga, A. C. Evans, P. T. Fox, and J. L. Lancaster. A probabilistic atlas of the human brain: Theory and rationale for its development. the international consortium for brain mapping (ICBM). *NeuroImage*, 2(2a):89–101, 1995.

[107] L. McEvoy, C. Fennema-Notestine, R. J.C., D. Hagler, D. Holland, D. Karow, C. Pung, J. Brewer, and A. Dale. Alzheimer disease: quantitative structural neuroimaging for detection and prediction of clinical and structural changes in mild cognitive impairment. *Radiology*, 251(1):1950–205, 2009.

[108] J. H. Morra, Z. Tu, L. G. Apostolova, A. E. Green, C. Avedissian, S. K. Madsen, N. Parikshak, X. Hua, A. W. Toga, C. R. J. Jr., M. W. Weiner, and P. M. Thompson. Validation of a fully automated 3D hippocampal segmentation method using subjects with Alzheimer's disease mild cognitive impairment, and elderly controls. *NeuroImage*, 43(1):59 – 68, 2008.

[109] J. H. Morra, Z. Tu, L. G. Apostolova, A. E. Green, C. Avedissian, S. K. Madsen, N. Parikshak, A. W. Toga, C. R. J. Jr, N. Schuff, M. W. Weiner, and P. M. Thompson. Automated mapping of hippocampal atrophy in 1-year repeat MRI data from 490 subjects with Alzheimer's disease, mild cognitive impairment, and elderly controls. *NeuroImage*, 45(1, Supplement 1):S3 – S15, 2009.

[110] J. Morris. The Clinical Dementia Rating (CDR): current version and scoring rules. *Neurology*, 43(11):2412 – 2414, 1993.

[111] S. G. Mueller, M. W. Weiner, L. J. Thal, R. C. Petersen, C. Jack, W. Jagust, J. Q. Trojanowski, A. W. Toga, and L. Beckett. The Alzheimer's Disease Neuroimaging Initiative. *Neuroimaging Clinics of North America*, 15(4):869 – 877, 2005.

[112] K. Niemann, A. Hammers, V. A. Coenen, A. Thron, and J. Klosterktter. Evidence of a smaller left hippocampus and left temporal horn in both patients with first episode schizophrenia and normal control subjects. *Psychiatry Research: Neuroimaging*, 99(2):93 – 110, 2000.

[113] K. M. Pohl, S. Bouix, and M. Nakamura et al. A hierarchical algorithm for MR brain image parcellation. *IEEE Transactions on Medical Imaging*, 26(9):1201–1212, 2007.

[114] O. Querbes, F. Aubry, J. Pariente, J.-A. Lotterie, J.-F. Dmonet, V. Duret, M. Puel, I. Berry, J.-C. Fort, and P. Celsis. Early diagnosis of alzheimer's disease using cortical thickness: impact of cognitive reserve. *Brain*, 132(8):2036–2047.

[115] E. M. Reiman, A. Uecker, R. J. Caselli, S. Lewis, D. Bandy, M. J. de Leon, S. D. Santi, A. Convit, D. Osborne, A. Weaver, and S. N. Thibodeau. Hippocampal volumes in cognitively normal persons at genetic risk for Alzheimer's disease. *Annals of Neurology*, 44(2):288–291, Aug. 1998.

[116] T. Rohlfing, D. B. Russakoff, and C. R. Maurer Jr. Performance-based classifier combination in atlas-based image segmentation using expectation-maximization parameter estimation. *IEEE Transactions on Medical Imaging*, 23(8):983–994, 2004.

[117] S. T. Roweis and L. K. Saul. Nonlinear dimensionality reduction by locally linear embedding. *Science*, 290(5500):2323–2326, 2000.

[118] D. Rueckert, L. I. Sonoda, C. Hayes, D. L. G. Hill, M. O. Leach, and D. J. Hawkes. Nonrigid registration using free-form deformations: Application to breast MR images. *IEEE Transactions on Medical Imaging*, 18(8):712–721, Aug. 1999.

[119] M. Sabuncu, B. Yeo, K. Van Leemput, B. Fischl, and P. Golland. A generative model for image segmentation based on label fusion. *IEEE Transactions on Medical Imaging*, 29(10):1714 –1729, 2010.

[120] M. R. Sabuncu, S. K. Balci, and P. Golland. Discovering modes of an image population through mixture modeling. In *MICCAI (2)*, volume 5242 of *Lecture Notes in Computer Science*, pages 381–389. Springer, 2008.

[121] J. W. Sammon. A nonlinear mapping for data structure analysis. *IEEE Trans. Computer*, 18(5):401–409, May 1969.

[122] B. Scherrer, F. Forbes, C. Garbay, and M. Dojat. Fully bayesian joint model for MR brain scan tissue, structure segmentation. In *MICCAI (1)*, Lecture Notes in Computer Science, pages 1066–1074, 2008.

[123] B. Schoelkopf, A. J. Smola, and K.-R. Mueller. Kernel principal component analysis. *Lecture Notes in Computer Science*, 1327:583–591, 1997.

[124] N. Schuff, N. Woerner, L. Boreta, T. Kornfield, L. M. Shaw, J. Q. Trojanowski, P. M. Thompson, J. C. R. Jack, M. W. Weiner, and the Alzheimer's; Disease Neuroimaging Initiative. MRI of hippocampal volume loss in early Alzheimer's disease in relation to ApoE genotype and biomarkers. *Brain*, 132(4):1067 – 1077, 2009.

[125] P. Shrout and J. Fleiss. Intraclass correlation: uses in assesing rater reliability. *Psychological Bulletin*, 86:420–428, 1979.

[126] J. G. Sled, A. P. Zijdenbos, and A. C. Evans. A nonparametric method for automatic correction of intensity nonuniformity in MRI data. *IEEE Transactions on Medical Imaging*, 17(1):87–97, Feb. 1998.

[127] S. Smith, N. D. Stefano, M. Jenkinson, and P. Matthews. SIENA – Normalised accurate measurement of longitudinal brain change. *NeuroImage*, 11(5, Supplement 1):S659 – S659, 2000.

[128] Z. Song, N. J. Tustison, B. B. Avants, and J. C. Gee. Integrated graph cuts for brain MRI segmentation. In *MICCAI (2)*, volume 4191 of *Lecture Notes in Computer Science*, pages 831–838. Springer, 2006.

[129] C. Studholme and V. Cardenas. A template free approach to volumetric spatial normalization of brain anatomy. *Pattern Recognition Letters*, 25(10):1191–1202, 2004.

[130] C. Studholme, D. L. G. Hill, and D. J. Hawkes. An overlap invariant entropy measure of 3D medical image alignment. *Pattern Recognition*, 32(1):71–86, 1999.

[131] J. Talairach, G. Szikla, P. Tournoux, A. Prossalentis, and M. Bornas-Ferrier. Atlas d'anatomie strotaxique du tlencphale. Etudes anatomo-radiologiques. 1967.

[132] S. Tang, Y. Fan, M. Kim, and D. Shen. RABBIT: rapid alignment of brains by building intermediate templates. In *SPIE*, volume 7259, 2009.

[133] J. B. Tenenbaum, V. de Silva, and J. C. Langford. A global geometric framework for nonlinear dimensionality reduction. *Science*, 290(5500):2319–2323, 2000.

[134] P. M. Thompson, K. M. Hayashi, G. I. de Zubicaray, A. L. Janke, S. E. Rose, J. Semple, M. S. Hong, D. H. Herman, D. Gravano, D. M. Doddrell, and A. W. Toga. Mapping hippocampal and ventricular change in Alzheimer disease. *NeuroImage*, 22(4):1754 – 1766, 2004.

[135] P. M. Thompson, R. P. Woods, M. S. Mega, and A. W. Toga. Mathematical/computational challenges in creating deformable and probabilistic atlases of the human brain, 2000.

[136] W. K. Thompson and D. Holland. Bias in tensor-based morphometry stat-roi measures may result in unrealistic power estimates. *NeuroImage*, In Press, Accepted Manuscript:–, 2011.

[137] T. H. Tomboy, A. Bar-Hillel, and D. Weinshall. Boosting margin based distance functions for clustering. 2004.

[138] J. Q. Trojanowski. Searching for the Biomarkers of Alzheimers. *Practical Neurology*, 3:30–34, 2004.

[139] M. Unser, A. Aldroubi, and M. Eden. Fast B-spline transforms for continuous image representation and interpolation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 13(3):277–285, Mar. 1991.

[140] F. van der Lijn, T. den Heijer, M. M. Breteler, and W. J. Niessen. Hippocampus segmentation in MR images using atlas registration, voxel classification, and graph cuts. *NeuroImage*, 43(4):708 – 720, 2008.

[141] L. van der Maaten, E. Postma, and J. van der Herik. Dimensionality reduction: A comparative review. *Published online*, 10:1–35, 2007.

[142] K. Van Leemput, F. Maes, D. Vandermeulen, and P. Suetens. Automated model-based tissue classification of MR images of the brain. *IEEE Transactions on Medical Imaging*, 18(10):897–908, 1999.

[143] E. M. van Rikxoort, I. Isgum, M. Staring, S. Klein, and B. van Ginneken. Adaptive local multi-atlas segmentation: application to heart segmentation in chest CT scans. volume 6914, pages 1–6. SPIE, 2008.

[144] P. Vemuri, J. L. Gunter, M. L. Senjem, J. L. Whitwell, K. Kantarci, D. S. Knopman, B. F. Boeve, R. C. Petersen, and C. R. J. Jr. Alzheimer's disease diagnosis in individual subjects using structural mr images: Validation studies. *NeuroImage*, 39(3):1186 – 1197, 2008.

[145] U. von Luxburg. A tutorial on spectral clustering. *Statistics and Computing*, 17(4):395–416, 2007.

[146] K. B. Walhovd, A. M. Fjell, J. Brewer, L. K. McEvoy, C. Fennema-Notestine, D. J. Hagler, R. G. Jennings, D. Karow, and A. M. Dale. Combining MR Imaging, Positron-Emission Tomography, and CSF Biomarkers in the Diagnosis and Prognosis of Alzheimer Disease. *American Journal of Neuroradiology*, 31(2):347–354+, 2010.

[147] P.-N. Wang, H.-C. Liu, J.-F. Lirng, K.-N. Lin, and Z.-A. Wu. Accelerated hippocampal atrophy rates in stable and progressive amnestic mild cognitive impairment. *Psychiatry Research: Neuroimaging*, 171(3):221 – 231, 2009.

[148] S. K. Warfield, K. H. Zou, and W. M. Wells III. Simultaneous truth and performance level estimation (STAPLE): an algorithm for the validation of image segmentation. *IEEE Transactions on Medical Imaging*, 23(7):903–921, 2004.

[149] G. Wenk. Neuropathologic changes in alzheimer's disease. *Journal of Clinical Psychiatry*, 64 Suppl 9, 2003.

[150] R. Wolz, P. Aljabar, J. Hajnal, and D. Rueckert. Manifold learning for biomarker discovery in MR imaging. In *Machine Learning in Medical Imaging*, volume 6357 of *Lecture Notes in Computer Science*, pages 116–123. Springer Berlin / Heidelberg, 2010.

[151] R. Wolz, P. Aljabar, J. V. Hajnal, A. Hammers, and D. Rueckert. LEAP: Learning embeddings for atlas propagation. *NeuroImage*, 49(2):1316 – 1325, 2010.

[152] R. Wolz, P. Aljabar, J. V. Hajnal, J. Lotjonen, and D. Rueckert. Manifold learning combining imaging with non-imaging information. In *IEEE International Symposium on Biomedical Imaging*, pages 1637–164, 2011.

[153] R. Wolz, P. Aljabar, R. A. Heckemann, A. Hammers, and D. Rueckert. Segmentation of subcortical structures and the hippocampus in brain MRI using graph-cuts and subject-specific a-priori information. In *IEEE International Symposium on Biomedical Imaging*, pages 470–473, 2009.

[154] R. Wolz, R. A. Heckemann, P. Aljabar, J. V. Hajnal, A. Hammers, J. Lotjonen, and D. Rueckert. Measurement of hippocampal atrophy using 4D graph-cut segmentation: Application to ADNI. *NeuroImage*, 52:1009 – 1018, 2010.

[155] M. Wu, C. Rosano, P. Lopez-Garcia, C. S. Carter, and H. J. Aizenstein. Optimum template selection for atlas-based segmentation. *NeuroImage*, 34(4):1612 – 1618, 2007.

[156] Z. Xue, D. Shen, and C. Davatzikos. Classic: Consistent longitudinal alignment and segmentation for serial image computing. *NeuroImage*, 30(2):388 – 399, 2006.

[157] D. L. Zhao, Z. C. Lin, R. Xiao, and X. Tang. Linear laplacian discrimination for feature extraction. In *Computer Vision and Pattern Recognition*, pages 1–7, 2007.

[158] L. Zöllei, E. G. Learned-Miller, W. E. L. Grimson, and W. M. Wells III. Efficient population registration of 3D data. In *CVBIA*, volume 3765 of *Lecture Notes in Computer Science*, pages 291–301. Springer, 2005.