Multi-Atlas Segmentation using Clustering, Local Non-Linear Manifold Embeddings and Target-Specific Templates

CHRISTOPH ARTHOFER, BSc., MSc.

Thesis submitted to the University of Nottingham
for the degree of Doctor of Philosophy

September 2017

Contents

Abstract **6**

Acknowledgement **7**

1 Introduction **8**

1.1 Medical image segmentation 8

1.2 Approaches to brain MR segmentation 11

1.2.1 General overview 11

1.2.1.1 Classical segmentation methods 12

1.2.1.2 Higher-level segmentation methods 13

1.2.2 Atlas-based segmentation 14

1.2.3 Multi-atlas segmentation (MAS) 18

1.3 Our approach 22

1.3.1 Overview of our approach 22

1.3.2 Contributions 24

1.3.3 Outline 25

2 Anatomical atlas building and MAS **27**

2.1 Introduction 27

2.1.1 Increasing the number of atlases 27

2.1.2 Joint group-wise registration and segmentation methods 28

2.1.3 Patch-based methods with linear registration 29

2.1.4 Reducing the computational burden 29

2.2 Pre-processing 30

2.2.1 Intensity inhomogeneity correction 30

2.2.2 Skull-stripping 31
2.2.3 Tissue classification 33
2.3 Image registration 35
2.3.1 Similarity measure 36
2.3.2 Transformation 37
2.3.3 Optimisation 40
2.4 Anatomical atlases and label probability maps 40
2.4.1 Single-subject atlases 41
2.4.2 Population atlases 42
2.4.3 Unbiased probabilistic atlas construction 45
2.4.4 Creation of label maps 48
2.5 Tools, datasets and evaluation strategies 50
2.6 Our approach to template building 53
2.6.1 Pre-processing 53
2.6.2 Estimating a population template 53
2.6.3 Transferring the label maps to the target 55
2.7 Experiments 56
2.8 Discussion 57

3 Target-specific template 63
3.1 Introduction 63
3.2 Comparison basis 67
3.2.1 Image intensities 67
3.2.2 Non-image information 67
3.2.3 Registration consistency 68
3.2.4 Anatomical geometry 69
3.2.5 Deformation fields 69
3.3 Similarity metrics 71
3.3.1 Basic similarity metrics 73
3.3.2 Manifold learning 75
3.3.2.1 Principal component analysis 76

3.3.2.2 Isomap 78
3.3.2.3 Local linear embedding 79
3.4 Our approach to TST building 80
3.4.1 Manifold embedding 81
3.4.2 TST construction 83
3.5 Experiments 84
3.5.1 Reconstruction of a GM image slice with eigenimages . 84
3.5.2 Reconstruction of a deformation field with eigenimages 86
3.5.3 Comparison of nonlinear dimensionality reduction methods and parameter optimisation 89
3.5.4 Method validation on the LONI dataset 92
3.5.5 Method validation on the ADNI-HarP dataset 94
3.5.6 Method validation on the IBSR dataset 96
3.5.7 Method validation on the NIREP-NA0 dataset 97
3.5.8 Method validation on the MICCAI 2012 dataset 98
3.5.9 Method validation on the MICCAI 2013 dataset 99
3.6 Discussion 100

4 Label Fusion **103**
4.1 Majority voting 104
4.2 Weighted voting 104
4.3 STAPLE 105
4.4 SIMPLE 107
4.5 STEPS 107
4.6 MALP 108
4.7 Joint label fusion 108
4.8 Patch-based label fusion 109
4.9 Our approach to propagation, fusion and binarisation 111
4.10 Experiments 113
4.10.1 Evaluation on the LONI dataset 113
4.10.2 Evaluation on the ADNI-HarP dataset 117

4.10.3 Evaluation on the IBSR dataset . . . 121
4.10.4 Evaluation on the NIREP-NA0 dataset . . . 122
4.10.5 Evaluation on the MICCAI 2012 dataset . . . 126
4.10.6 Evaluation on the MICCAI 2013 dataset . . . 127
4.11 Discussion . . . 130

5 Dynamically adjusted labels **134**
5.1 Offline learning . . . 134
5.2 ROI selection . . . 136
5.3 Clustering methods . . . 137
5.3.1 K-means . . . 138
5.3.2 Affinity propagation . . . 138
5.3.3 Hierarchical agglomerative clustering . . . 139
5.4 Our approach to label adjustment . . . 141
5.4.1 Dividing the labels into clusters . . . 141
5.4.2 Combining the clusters . . . 142
5.5 Experiments . . . 143
5.5.1 Evaluation on the ADNI-HarP dataset . . . 143
5.5.2 Evaluation on the MICCAI 2013 dataset . . . 146
5.5.3 Evaluation on the NIREP-NA0 dataset . . . 146
5.6 Discussion . . . 148

6 Application to Tourette's images **152**
6.1 Introduction . . . 152
6.1.1 Healthy brain development . . . 152
6.1.2 Tic disorders and Tourette's Syndrome . . . 154
6.1.3 Method . . . 155
6.2 Results . . . 156
6.3 Discussion . . . 156

7 Conclusion and perspectives **161**
7.1 Chapter overview . . . 161

7.2 Conclusion 166
7.3 Perspectives 167

Bibliography **169**

Abstract

Multi-atlas segmentation (MAS) has become an established technique for the automated delineation of anatomical structures. The often manually annotated labels from each of multiple pre-segmented images (atlases) are typically transferred to a target through the spatial mapping of corresponding structures of interest. The mapping can be estimated by pairwise registration between each atlas and the target or by creating an intermediate population template for spatial normalisation of atlases and targets. The former is done at runtime which is computationally expensive but provides high accuracy. In the latter approach the template can be constructed from the atlases offline requiring only one registration to the target at runtime. Although this is computationally more efficient, the composition of deformation fields can lead to decreased accuracy.

Our goal was to develop a MAS method which was both efficient and accurate. In our approach we create a target-specific template (TST) which has a high similarity to the target and serves as intermediate step to increase registration accuracy. The TST is constructed from the atlas images that are most similar to the target. These images are determined in low-dimensional manifold spaces on the basis of deformation fields in local regions of interest. We also introduce a clustering approach to divide atlas labels into meaningful sub-regions of interest and increase local specificity for TST construction and label fusion. Our approach was tested on a variety of MR brain datasets and applied to an in-house dataset.

We achieve state-of-the-art accuracy while being computationally much more efficient than competing methods. This efficiency opens the door to the use of larger sets of atlases which could lead to further improvement in segmentation accuracy.

Acknowledgement

Throughout the journey leading to this Ph.D. thesis I have been supported in various ways and influenced by many people, making it a difficult task to acknowledge each and every single one in this section.

First and foremost, I would like to thank my primary supervisor Dr. Alain Pitiot for the countless meetings and inspiring discussions, the constructive feedback and comments, and the encouragement and motivation, all of which will be missed. I would also like to send my gratitude to my second supervisor Prof. Paul Morgan for sharing his extensive scientific experience and giving valuable advice. I am very grateful for having had such kind and thoughtful supervisors. Working as an MR scanner operator alongside my Ph.D. provided complementary experience from the MR acquisition side and was made possible by Jan Alappadan Paul and Prof. Penny Gowland. Thank you also to Prof. Stephen Jackson and the James Tudor Foundation for the generous funding support and the opportunity to work in his laboratory. I hope we will cross paths again in the future. I would also like to thank my examiners Dr. Juan Eugenio Iglesias and Prof. Tony Pridmore for their time and helpful feedback.

I would like to thank my loving family, Elisabeth, Rainer, Christine, Karl and Berta for their encouragement and support despite being 1204 km away. I will be forever grateful for their continued investment and confidence in me achieving my dreams. If it was not for them this Ph.D. would not have been possible. I am also very blessed to have Filipa in my life, who has spread optimism and given me strength along the way.

Thank you to Darren, Antonis and Tom, who filled the office with a great atmosphere and to Susi, Winti, Kathi, Martina, Sebastian, Niki, Johanna and Benjamin for their unwavering friendship.

Chapter 1

Introduction

Image segmentation is the process of partitioning an image into multiple components. While manual segmentation is time-consuming and poorly reproducible, automated methods, and in particular atlas-based segmentation, have shown to be efficient alternatives. In this chapter the aims of image segmentation and various different approaches for the segmentation of MR images will be presented. Due to our main interest in atlas-based segmentation methods, we will provide an overview of its basic underlying concept as well as advanced solutions followed by an outline of our approach and contributions.

1.1 Medical image segmentation

Aims of medical image segmentation: Image segmentation, in the general sense, is defined as the partition of an image into nonoverlapping homogeneous regions or objects, with respect to certain features or characteristics, and their separation from the background [166]. Segmentation, applied in the field of medicine and, in particular, to medical images acquired by magnetic resonance imaging (MRI), is commonly required for delineating anatomical regions of interest such as the brain, different tissue types such as grey matter or white matter, or regions associated with function, activity

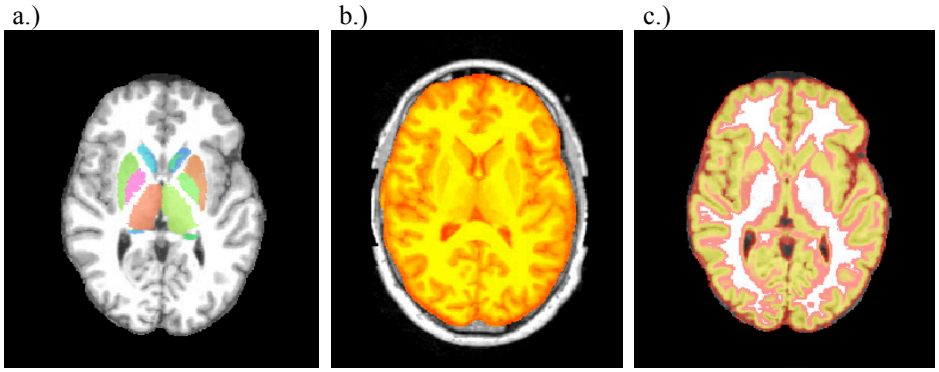


Figure 1.1: Aims of medical image segmentation on examples of segmented a) regions of interest, b) brain and c) grey matter.

or pathology (Fig. 1.1). It is often one of the first steps in a series of image analysis tasks such as planning of medical treatment or surgery [61, 143], diagnosis and patient follow-ups [114], and monitoring of disease progression or development [34]. Consequently, we require segmentation methods that provide accurate and reproducible results.

Challenges in image segmentation: Typically, the segmentation of an image combines two main challenges. The first is to identify certain image features like homogeneity, continuity or similarity extracted from intensity values, differences in texture or gradients in border areas (Fig. 1.2). The second challenge is to transform these extracted features into semantic entities and provide context. This makes it a classification task, which is one of the most difficult tasks in the image processing domain [28, 91, 210]. The segmentation of an image yields groups of voxels that belong to the same structure or region of interest, requiring each voxel of the image to be classified and assigned to one of the groups. Although, over the years, various concepts have been extensively studied, the increasing number of different requirements and their associated challenges allow for no general solution which is applicable to problems from all disciplines [76].

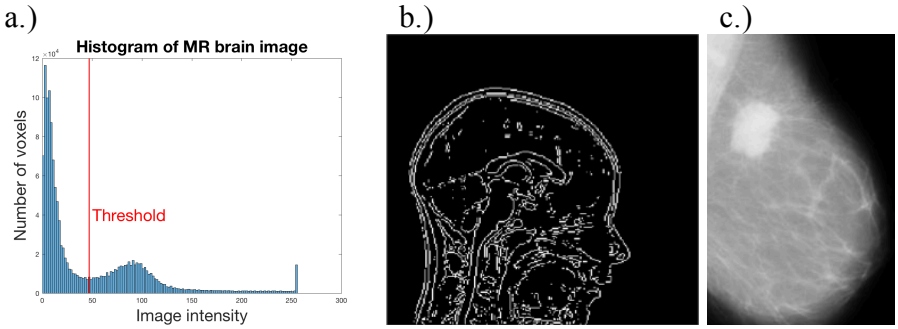


Figure 1.2: Segmentation based on a.) image intensities [28], b.) gradients and c.) texture [116].

Common problems in MRI: The acquisition of images with different modalities, scan protocols or with scanners from different manufacturers, as well as the underlying physical signal acquisition itself pose challenges such as [232]:

- intensity non-uniformities, e.g. bias fields where voxels of varying intensities belong to the same tissue type,
- movement artefacts known as ghosting,
- partial volume effects where voxel intensities represent a mixture of different tissue types,
- frequency/phase wrapping caused by aliasing,
- noise,
- poor image contrast and
- weak boundaries.

All of these artefacts can have an impact of varying degree on the quality of the image processing pipeline. This makes early detection and, where possible, correction crucial.

1.2 Approaches to brain MR segmentation

1.2.1 General overview

Although manual segmentation and interpretation by experts is still considered as the gold standard, it does not come without drawbacks. It is a very time-consuming and cost-intensive task, since every voxel has to be assigned a label. The segmentation accuracy is restricted by the variability introduced due to the operator(s), as different experts produce different segmentation results of the same structure (inter-observer variability). Even when the same expert performs the segmentation of an image repeatedly, variability between the resulting segmentations is introduced (intra-observer variability) [232]. An early assessment by Clarke *et al.* [53] outlined validation methods to measure reproducibility and compared various volume measurement studies. For example, the segmentation of the hippocampus with manually supervised methods showed inter- and intra-observer variability of 14% and 10% respectively. Although technical equipment for acquisition and screening of MRI and segmentation protocols such as the Harmonized Hippocampal Protocol [35, 84] have improved since Clarke *et al.'s* study approximately 20 years ago, manual and manually supervised segmentation is still used as the gold-standard. An evaluation of the commonly used public dataset from the more recent MICCAI 2013-SATA challenge showed inter-scan reliability of 68% for the basal forebrain and 79% for the middle occipital gyrus, measured with the the Dice overlap coefficient [17].

In order to improve accuracy with respect to a manual gold-standard and reproducibility, a wide variety of automated segmentation algorithms have been developed. They can be classified in many ways [166] such as based on their degree of automatism (manual, semiautomatic, automatic), their spatial extent (local pixel-based, global region-based), or the concept (area-based, edge-based). Methods can also be categorised into classical (thresholding, edge-based, region-based), statistical and neural network techniques [160]. In the following sections we will give an overview of commonly used classical

and higher-level methods. In general, the former are based on low-level image processing, while the latter use classification and clustering methods which can also incorporate a priori anatomical knowledge.

1.2.1.1 Classical segmentation methods

One fast and simple method is global binary ***thresholding*** where a fixed threshold is used to split the image based on its pixels' intensity values. Support for finding the ideal threshold is given by analysing the histogram of the entire image [120, 149, 164]. This approach was further extended for multilevel thresholding [242], allowing the original image to be divided into more than two classes and local thresholding [48] for more regional decision making. In general, thresholding is simple to implement and yields fast computation times, but is influenced by the amount of noise, intensity contrast and anatomical complexity in the image.

In contrast, ***region growing*** aims to merge pixels into homogeneous, connected regions [2]. Every pixel in the neighbourhood around one or more starting points is examined and added to the region if a common homogeneity criterion is fulfilled. The outcome of the algorithm strongly depends on the chosen condition and is heavily influenced by the defined starting points. Conversely, instead of growing a region by merging pixels, one region can be divided into sub-regions based on a homogeneity criterion until no further splitting is possible [39]. In order to combine the advantages of both methods, a split-and-merge algorithm [103] was developed.

Another computationally efficient group of methods is based on determining the edges of a structure. Commonly used ***edge detectors*** are the Sobel [192], Prewitt [158] or Canny [42] gradient operators based on first order derivatives, or the Laplacian operator based on second order derivatives. Due to the operators' sensitivity to noise, smoothing the image is recommended.

The ***watershed*** algorithm [32] considers the 2D grayscale gradient or topographic distance image of the target image as a 3D surface. The grayscale

values represent heights with local minima as sources of basins and increasing values as ridges. Starting at each source, the basins are flooded. As the water rises and water sources meet, a barrier, i.e. a watershed, is created. The method tends to over-segment images, which usually requires merging of partitions in a post-processing step.

1.2.1.2 Higher-level segmentation methods

Clustering methods such as ***unsupervised methods*** aim at partitioning the data into non-overlapping groups based on certain homogeneity attributes as measured by the clustering criterion. Since each clustering technique implicitly imposes a structure on the data, the method should be chosen carefully by considering the data under analysis. Commonly used approaches include hierarchical methods and methods based on a sum of square error criterion such as k-means [137]. Clustering methods will be discussed in Section 5.3 in more detail.

Alternatively, image segmentation can be seen as a pixel classification problem. ***Statistical approaches*** aim at characterising a structure and assigning it a category based on its features or attributes. Classifiers belong to the group of ***supervised methods***, which discriminate new input data based on a classifier learned from a pre-labelled training set. Examples for supervised segmentation methods include ***active shape*** [63] and ***active appearance models*** [62], which are based on active contour models [54, 115] and use deforming curves or surfaces to segment a structure. These curves are controlled by internal and external forces where external forces pull the curve towards a desired object shape and internal forces preserve the local tension or smoothness of the curve.

Due to the availability of large annotated datasets ***neural networks (NNs)*** have gained a lot of attention. NNs consist of a large number of interconnected nodes, or neurons. Each neuron has a set of incoming connections and one outgoing connection, which represent weighted inputs and the output respectively. By organizing these networks into layers, the weights

and topological relationships between the variables can be updated and dynamically adjust to a task, which allows the modelling of complex nonlinear relationships. Networks with multiple layers between input and output are considered as ***deep neural networks***. One type of commonly used NNs are ***convolutional neural networks (CNNs)*** where the input image is convolved with kernels at every layer. The convolutions generate sets of features which are non-linearly transformed. CNNs usually include pooling layers where the output of previous layers is combined by applying filters. CNNs can classify each pixel individually or produce likelihood maps.

Another important and powerful high-level technique which incorporates anatomical a priori information is ***atlas-based segmentation*** [170]. In the following sections we will introduce single- and multi-atlas segmentation before presenting an overview of our approach.

1.2.2 Atlas-based segmentation

The traditional concept of atlas-based segmentation uses a single model of the human brain, which will be referred to as an atlas, to segment a new anatomical intensity brain image, referred to as the target (Fig. 1.3). The atlas, in its simplest form, consists of one individual anatomical intensity image and its corresponding label map (see Section 2.4.1). It is of importance to note that in the literature the term atlas is sometimes not further specified and might refer to an individual single atlas or a probabilistic atlas. In this manuscript special care will be taken to indicate the type of atlas if not evident from the context. The target image is aligned to the anatomical atlas image by estimating the spatial transformation between them using image registration. This process aims at finding the best mapping between the anatomical structures in the target and the corresponding anatomical structures in the atlas intensity image. The label maps from the atlas are transferred to the target by applying the inverse of the same transformation

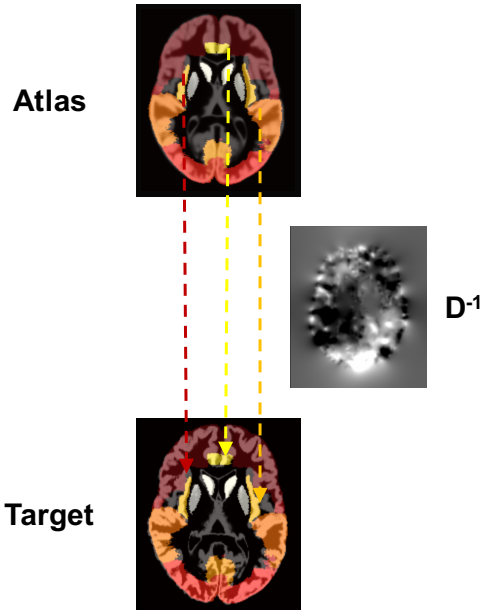


Figure 1.3: Single atlas-based segmentation concept.

to them. This algorithm crucially depends on the quality of the registration, which is the reason why in the literature it is often referred to as a registration problem rather than a segmentation problem. One of the first implementations of this basic concept for 3-D information propagation was presented by Collins *et al.* [56] and Dawant *et al.* [69]. The advantage of their concept is its independence from the selected atlas labels, which allows the use of multiple different atlas label maps without the need to re-compute the mapping between atlas- and target-space. The registration can be performed with a time-efficient affine registration or with a nonlinear method. In general a linear registration leads, without further post-processing, to a poor alignment, and, in turn, poor segmentation quality. Commonly a combination of linear and computationally more expensive nonlinear registration methods is used, which leads to a better alignment of anatomical structures and consequently better segmentation results. For instance, Christensen *et al.* used a diffeomorphic registration method [50], which used a low-dimensional transformation for global shape differences and a high-dimensional vector field for the alignment of fine structures based on linear elasticity and fluidity models.

Their atlas-based model was used to automatically label a cortical surface by elastically deforming a pre-labelled 3-D surface atlas so that the cortical sulci of the atlas align with the corresponding image features in the target image [68, 179].

The segmentation accuracy crucially depends on the quality of the mapping between the target and atlas images. A one-to-one mapping between every point in the two images might not always be possible due to the high anatomical variability. Over the last decades, anatomical variability has been investigated in the whole brain as well as for particular structures showing that volume, distribution of grey and white matter, and anatomical shape vary considerably between individuals [9, 24, 152, 231]. The quality of the registration is also limited by the registration model and can lead to registration errors (see Section 2.3). In an extensive evaluation, one linear and 14 nonlinear registration methods have been compared on a set of 80 manually labeled brain images [122]. One of the surrogate evaluation strategies measured the overlap as the agreement of deformed source and target label volumes averaged across all regions and brain pairs. A maximum overlap of approximately 71% was reached for the LONI-LPBA40 dataset [185] with the SyN registration method [21]. One way to improve registration accuracy is to use an atlas which is expected to lead to the most precise registration and, in turn, segmentation. The comparison of different atlas selection strategies has shown that an individual atlas with an intensity image more similar to the target can achieve better segmentation accuracy than the use of a randomly selected atlas [57, 85, 169]. The choice of atlas also depends significantly on the ROI [72]. While a single atlas might provide the best possible result for some ROIs it does not necessarily lead to the best outcome for other ROIs. It was concluded that there is no single atlas obtained from one individual subject that would provide the overall most accurate segmentation of a new target image. Consequently, it has been suggested to use more than one atlas to capture a wider range of inter-subject variability. In a series of studies

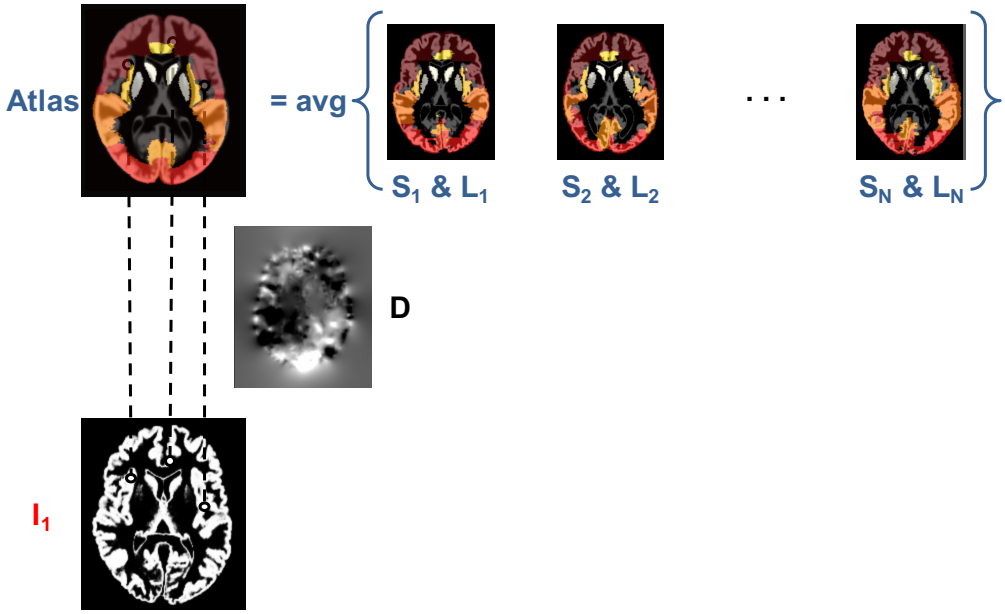


Figure 1.4: Probabilistic atlas-based segmentation concept.

the use of multiple individual atlases over a single individual atlas has shown improved segmentation accuracy [4, 169, 239].

One way to make use of multiple atlases is, as previously indicated, the selection of the single most similar atlas to the target from a set of atlases and its use for segmentation. Although this strategy accesses the information of multiple atlases, it eventually does not fully take advantage of all atlases.

Another way to incorporate the labelling information of multiple atlases is by building an average-, population-, statistical- or probabilistic atlas (Fig. 1.4) which provides a value for each voxel indicating its probability of being part of a particular label (see Sections 2.4.2 - 2.4.3). Similar to the procedure with an individual atlas, the probabilistic atlas intensity image is registered to the target and its corresponding probabilistic label maps propagated. The use of one standardised space presents a large advantage in terms of performance. It requires only one nonlinear registration to the target at runtime and the impact of registration errors can be partly alleviated by the probabilistic maps. All individual atlases and registered targets are linked via their

deformation fields which makes the mapping between each of them possible by composing the respective deformation fields. In comparison, without an average atlas, computationally expensive nonlinear registrations have to be performed between each of the individual atlases and every new target.

Although computationally very efficient, the use of a probabilistic atlas has shown to provide less accurate segmentation results compared to the direct use of each individual atlas for target segmentation, which will be referred to as multi-atlas segmentation and explained in more detail in the next section.

1.2.3 Multi-atlas segmentation (MAS)

In the previous section, segmentation based on a single atlas was introduced. The single atlas of choice can either be from one individual, which, for example, could be selected as the best possible match from a set of individual atlases, or be represented as a probabilistic atlas, constructed from multiple individual atlases.

In contrast to a probabilistic atlas, Multi-Atlas Segmentation (MAS) directly utilises the label information from all individual atlases of a set. A new target image is nonlinearly registered to each individual atlas resulting in as many deformation fields as there are atlases (Fig. 1.5). The same deformation fields can be applied to the corresponding atlas label maps to warp them to the target, resulting in candidate labels, which can be combined into the final segmentation with a label fusion algorithm. This was shown to outperform single atlas-based segmentation methods [169] and is commonly used for segmentation problems. This is in line with the findings in the pattern recognition literature where the combination of classifiers is in general more accurate than one individual classifier [38].

One of the first implementations was proposed in a series of papers by Rohlfing *et al.* [169, 168, 172, 173]. In their most influential work [169] they used an optimised free-form deformation algorithm to non-rigidly register

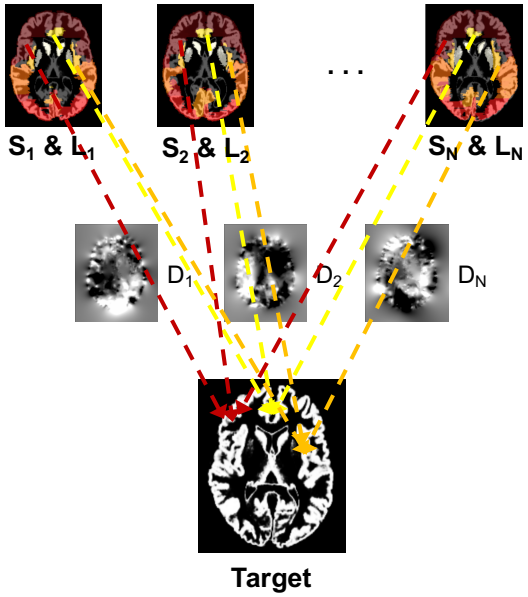


Figure 1.5: Multi atlas-based segmentation concept. Each individual atlas intensity image is registered to the target and each resulting deformation field is used to transform the corresponding label map to the target.

each of the atlas images to the remaining images yielding one segmentation candidate from each. The final segmentation was determined by majority voting, which assigns to each voxel the label with the most occurrences at the corresponding location in the candidate images. In [172] this simple vote rule or majority voting for the fusion of candidate labels was further extended with the expectation maximisation label fusion algorithm by Warfield *et al.* [228], originally designed for the assessment of human labelling performance and estimation of the true segmentation. Since most of Rohlfing's initial work was done on bee brains, more papers about atlas selection strategies corroborated his findings for automatic labelling of the human brain [23, 123, 239] and the impact of the label fusion method was investigated by Heckeman *et al.* [98]. However, most top performing MAS methods rely on the accurate alignment between each atlas and the target, which is usually done by estimating the pairwise nonlinear spatial correspondence between each atlas and the target at runtime. This step is typically the most time-consuming part of the MAS

algorithm.

Some methods have focused on running speed, e.g. Coupe *et al.*'s patch-based approach [64] requires only linearly aligned atlases but achieved the lowest performance in Wu *et al.*'s evaluation [237]. Two alternative strategies for reducing the number of registrations and improving registration accuracy either employ a target-specific template (TST), more similar to the target than an average population template [59, 61, 151, 187, 235], or construct a graph structure of intermediate similar atlas images [46, 88, 96, 110, 118, 147, 202, 225, 234] to split a large, potentially imprecise registration, into smaller more accurate deformations. However, TSTs are usually constructed as probabilistic atlases which does not allow the direct consultation of the warped individual atlas label maps. In most methods these probabilistic atlases are constructed from the locally or globally most similar individual atlas intensity images to the target by registering them iteratively to their average [59], which can increase the computational burden at runtime. Similarity measures such as normalised mutual information or the sum of squared differences are commonly used for comparison of intensity images. More recent methods have shown improved accuracy for candidate selection and label fusion by using distances between images calculated in manifold space [75]. Decisions solely based on image intensity values could potentially be corrputed by artefacts or inhomogeneities. In the presence of anatomical pathology the use of deformation fields for similarity comparison is more robust [59, 61, 151, 163]. Other methods split the given labels into smaller, more localised sub-regions with the goal to improve candidate selection and fusion [18, 126, 217, 178]. However, most methods split the regions randomly without taking contextually meaningful information into account. One drawback of graph-based strategies is the observed decrease in segmentation overlap almost linear to the number of composed links between intermediate templates [188].

In general, eight commonly found components in MAS methods (Fig. 1.6) were identified by Iglesias and Sabuncu [105], with some of them being op-

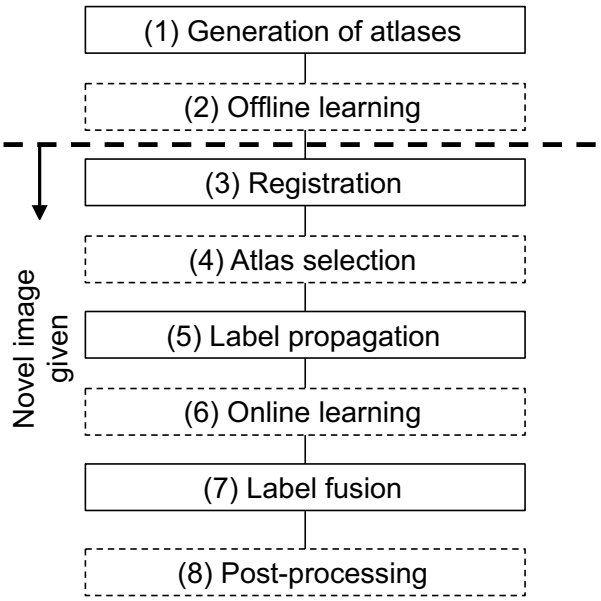


Figure 1.6: Components of MAS. Optional components are indicated by the dashed boxes. Figure from Iglesias and Sabuncu [105]

tional. Given a set of atlases (1), some methods learn from or analyse the atlases offline, for example by constructing various templates or a graph structure (2). At runtime, this information is ready to use, which, in turn, improves computational efficiency. Once a target image is given, one or more atlas images are registered to it (3). Based on similarity criterions the atlas(es) with the highest expected probability of segmenting the target accurately can be selected (4) and the labels propagated (5). At runtime, another learning step can be performed (6) which can provide additional information for the concurrent label fusion (7). Some algorithms additionally apply post-processing methods (8) to smooth the borders or use the fused labels to initialise an active contour or level set algorithm. Over the years each of these areas has become a topic of research. In this thesis, we experimented with many approaches in some of those areas, and developed our own. The next section provides an outline of our method and our main contributions.

1.3 Our approach

1.3.1 Overview of our approach

Motivated by the necessity to segment brain structures in MRI scans both accurately and in a time efficient manner, we have developed a novel MAS approach with the main aims of (1) reducing the number of nonlinear registrations at runtime, and (2) providing state-of-the-art accuracy and (3) robust results on diverse datasets.

To reduce the number of registrations at runtime we use both an average population template and a TST. For clarification we introduce the following notation: a ROI refers to the anatomy of interest, for example the hippocampus; a label refers to the particular delineation of a ROI, such as the hippocampus as annotated in one atlas; region refers to a collection of voxels in images or deformation fields without regards to a particular ROI; cluster refers to one sub-division of a region with similar attributes.

Offline, we construct a population template from a set of pre-segmented, affinely-registered atlas images and, by the same token, estimate a non-linear deformation field D between each atlas image $\widetilde{I}$ and the template $\overline{G}$ (Fig. 1.7.a). The same deformation fields are used to transform the corresponding atlas labels $\widetilde{L}$ to $\overline{G}$ and to find clusters that undergo similar or different deformation. Firstly, for each location we calculate the standard deviation of the corresponding displacement magnitudes of the deformation fields. We assume that a high standard deviation indicates dissimilar deformation within the dataset at this location, while a low standard deviation indicates a location with similar displacement in the dataset. Secondly, we calculate the union from the labels in the space of $\overline{G}$, which provides the largest region covered by the labels of a ROI. For each union we cluster the corresponding voxels based on their standard deviations of the deformation fields and locations. The resulting set of clusters are warped back on each of the atlases, where we use the atlas labels to crop them.

At runtime, we non-linearly register the previously unseen target image

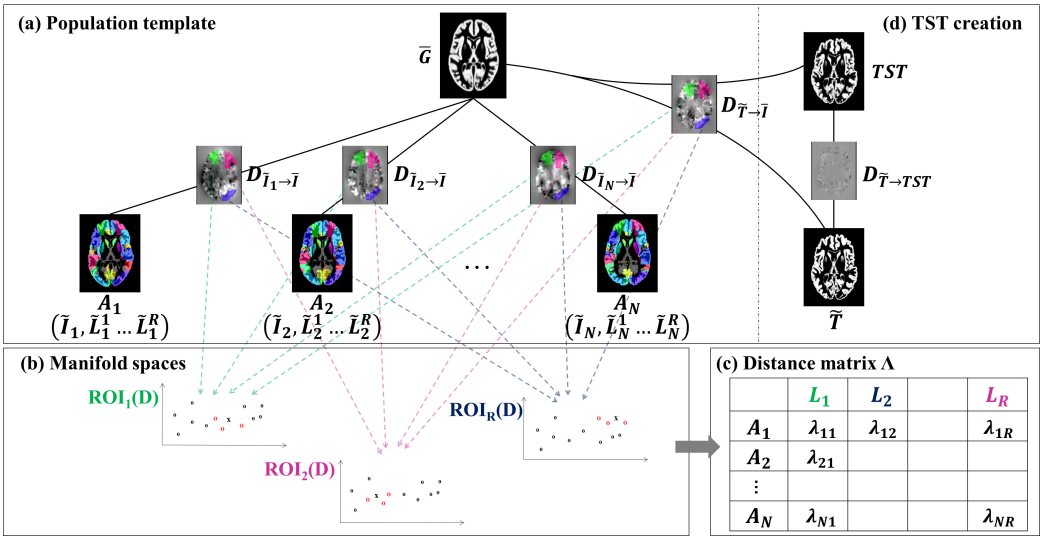


Figure 1.7: Overview of our approach. The population template $\overline{G}$ is constructed from individual atlases offline (a). At runtime the target $\widetilde{T}$ is registered to $\overline{G}$. A manifold embedding is constructed from the deformation fields for each ROI (b) allowing the ranking of the atlases based on their similarity to the target (c). A TST is constructed from the most similar atlas images warped on $\widetilde{T}$ by composing $D_{\widetilde{I}_j \to \overline{I}}$ and $D_{\widetilde{T} \to \overline{I}}$ (d). $\widetilde{T}$ is registered to the TST and the atlas label maps are propagated to $\widetilde{T}$ by composing $D_{\widetilde{I}_j \to \overline{I}}$, $D_{\widetilde{T} \to \overline{I}}$ and $D_{\widetilde{T} \to TST}$ where they are fused.

$\widetilde{T}$ to the template to obtain the target deformation field (first runtime registration). For each cluster we then create a manifold embedding from the corresponding regions extracted from the atlas and target deformation fields (Figure 1.7.b). The atlas regions are then ranked in order of similarity to the target region using their L^2 distances in manifold space (Fig. 1.7.c). For each label the atlas images are then warped onto the target and a TST is constructed, to which the target is non-linearly registered (second runtime registration). We then project the label maps onto the target by composing the deformation fields between the atlases and the population template, between the population template and the TST, and between the TST and the target (Fig. 1.7.d). The resulting candidate segmentations are fused using

weights based on the rankings in manifold space which allows to determine areas with low and high probabilities. In addition to the probability values, the final decision whether a voxel is part of a label is also based on the corresponding intensity value of the target by comparing it to the intensity distribution determined from high-probability areas.

1.3.2 Contributions

In order to combine the efficiency of single-atlas based methods and the accuracy of MAS methods we propose a novel framework with the following improvements.

First contribution: Minimising the number of registrations without compromising accuracy

(1) We reduce the number of nonlinear registrations at runtime by constructing an average population template from all atlas images offline (see Chapter 2) and a TST from the most similar atlas images to the target online, requiring only two nonlinear registrations at runtime for each new target.

(2) In contrast to other methods, the atlas images are non-linearly aligned with the target in an efficient way allowing the construction of the TST from the warped atlases on the target. This makes for an even more similar TST to the target than an average atlas or a target-specific probabilistic template constructed in average atlas space.

(3) The target is non-linearly registered to the morphologically very similar TST and by composing the deformation fields, all atlas label maps can be propagated directly to the target and consulted for fusion.

Second contribution: TST construction using region-specific manifold embeddings of deformation fields

(4) We create a nonlinear manifold for each ROI rather than from the whole image and from the deformation fields rather than from the intensity images

to provide the most accurate weights for the TST construction and label fusion (see Chapters 3 - 4). We evaluated two nonlinear manifold embedding strategies on a dataset with over 50 ROIs and used a more region-independent fusion method.

Third contribution: Splitting labels into meaningful clusters
5) In order to automatically select relevant sub-regions to allow for more accurate weight estimation and concurrent fusion our last contribution is the clustering of pre-defined ROIs based on the variability of the deformation. This allows the characterisation of regions that require a high and low amount of deformation at the population level (see Chapter 5).

1.3.3 Outline

In the remainder of this thesis we will discuss relevant methods from each of the components of the MAS concept (Fig. 1.6) and the impact of our approach and contributions. In the beginning of each chapter we will provide a literature overview followed by a presentation of our own approach and experiments.

Chapter 2 outlines strategies for the spatial alignment of atlases with the target. We elaborate on the use of a population atlas for MAS and techniques for its creation.
Chapter 3 provides an overview of computationally efficient MAS solutions. We present a fast and accurate MAS strategy that employs intermediate templates and manifold learning.
Chapter 4 introduces and compares techniques for the fusion of candidate labels and the thresholding of probabilistic segmentations.
Chapter 5 outlines strategies for offline learning to reduce the computational burden at runtime and/or improve segmentation accuracy. We present our approach of using localised sub-regions which are dynamically derived

from the given atlas labels.

In ***Chapter 6*** we present the segmentation results obtained by applying our method to brain MR scans of healthy subjects and Tourette's patients.

Chapter 2

Anatomical atlas building and multi-atlas segmentation

2.1 Introduction

One of the essential parts of the MAS concept is the spatial alignment of the atlases with the target, which is achieved with image registration. In this chapter we provide an introduction to image registration methods, their use in MAS approaches and their effect on segmentation accuracy and runtime. In the remainder of this section we provide an overview of solutions for registration in MAS and commonly used pre-processing steps, before outlining image registration methods in more detail, and techniques to create atlases and construct population templates.

2.1.1 Increasing the number of atlases

The quality of an image registration method depends on various factors such as the characteristic of the image, the ROI to be segmented and the underlying registration algorithm and its parameters. Different methods or even the same method repeatedly applied with different parameters yield different outcomes. Similarly to the pattern recognition literature, where it has been shown that multiple classifiers yield improved and more stable results

over single classifiers [38], Rohlfing and Maurer [171] utilised these different outcomes to create a larger set of atlas-based classifiers. They repeatedly applied the same registration method with different sets of parameters to the same image, with the goal of improving classification accuracy. In related work by Doshi et al. [74], multiple different image registration methods were applied to the same images, leading to a set of multiple different registration results. A larger set of classifiers increased the probability of finding suitable candidates and showed superior segmentation results over using a single method. However, while a larger set of different classifiers can improve classification accuracy, the larger number of pairwise registrations with one or even multiple methods represents the most time-consuming part of MAS.

2.1.2 Joint group-wise registration and segmentation methods

Traditionally the transformation is estimated between each individual atlas and the target, independently from the registrations of other atlases to the target. Groupwise methods [3, 97, 106, 110, 223] consider the strong reliance of the segmentation outcome on the quality of the registration and, conversely, the potential improvement in registration quality by incorporating label information, i.e. a more precise deformation field for label propagation provides better segmentation results and atlas labels provide important features to improve registration accuracy. For example, the generative probabilistic model by Iglesias, Sabuncu and Van Leemput [106] uses the consistency of voxel intensities within an ROI to simultaneously estimate the atlas-to-target deformation fields and target labels. In contrast to other MAS methods, these deformation fields are considered as model parameters alongside the Gaussian distribution parameters of the voxel intensities. The most likely values are estimated by a variational expectation algorithm. Although it can increase registration and segmentation accuracy, it can also have a negative impact on computational efficiency if all of the steps of the

method have to be re-applied for every new target image.

2.1.3 Patch-based methods with linear registration

One class of MAS methods which was originally proposed with only linearly aligned atlases is based on the non-local means method [41]. It was first utilised in MAS by Coupe et al. [64] and uses patches for the classification of target voxels (see Section 4.8). Patches around each voxel and around its neighbours in the atlas images are compared to the corresponding patch in the target image. Based on their similarity and the atlas labels at the specific location, a label can be assigned to the target voxel. Patches are selected within a specified search neighbourhood around each voxel which relaxes the one-to-one correspondences between target and atlases and makes the alignment of the images less constrained. Although proposed without the need for time-consuming nonlinear registrations, the repeated search through all of the voxels can still take a considerable amount of time and achieved less accurate results compared to other patch-based methods using nonlinear registrations. This lead to various approaches combining both nonlinear registration- and patch-based characteristics, which showed better results than the individual methods [13, 20, 25, 81, 174, 181].

2.1.4 Reducing the computational burden

The use of pairwise registrations, multiple pairwise registrations per atlas, group-wise registration, or patch-based schemes provides improved segmentation results at the cost of computational efficiency at runtime.

This computational burden at runtime can be reduced by using a common coordinate system to spatially normalise all atlases offline. At runtime, a new target requires only one registration to this same space. In the literature three main approaches for the selection of the coordinate system can be identified. Firstly, a standardised individual brain image that is not part of the atlas set can be used for normalisation, e.g. Aljabar et al. [4, 5].

Secondly, one of the atlas images can be selected as a reference space [188, 217]. However, the selection of a standardised brain image or an atlas image from the cohort might differ substantially from the remaining atlases or the target and might lead to inaccurate registration results. The third approach facilitates a population template, constructed from all atlases offline [12, 60, 61, 70, 81, 161, 187]. Although it does not guarantee similarity to the target, this approach can capture the variability within the atlas population, and, consequently, provide more accurate alignment in the average reference space.

2.2 Pre-processing

2.2.1 Intensity inhomogeneity correction

A general problem in MRI is the smooth intensity variation that can occur across the whole image. It can be caused by static field inhomogeneity and RF coil nonuniformity, and depends on pulse sequence and field strength. In most applications the correction for inhomogeneities is performed as a post-processing step by image processing tools such as SPM or Freesurfer [67] in addition to hardware-related solutions in MR acquisition. In SPM [15] inhomogeneity correction is automatically performed when segmenting an image into grey matter (GM), white matter (WM) and cerebrospinal fluid (CSF). The tissue classification in SPM is based on prior probability maps and a mixture of Gaussians model with more than one Gaussian for each class. The objective function of the model is extended by extra parameters to take intensity inhomogeneity into account. It uses a linear combination of low-frequency sinusoidal basis functions to model the spatially smooth nonuniformity. This integrated approach has shown to outperform competing methods [87] and can be applied to images acquired with different pulse sequences. However, the integration of the correction method into the segmentation algorithm which, in turn, is based on prior knowledge about tissue types, does not allow its independent application to other anatomical struc-

tures than the brain.

A dedicated method for intensity inhomogeneity correction is nonparametric nonuniform intensity normalisation (N3) [190]. Although it is conceptionally similar to SPM [130], N3 works without tissue class models, can be applied to pathological data, and is independent of the pulse sequence. The inhomogeneous field is assumed to blur the image which consequently reduces the high frequency components of the intensity distribution. The objective can be stated as finding the smooth, slowly varying, multiplicative field that maximises the high frequency content of the unblurred intensity distribution. This is achieved by iteratively sharpening the intensity distribution of the blurred image and estimating the smooth field, which produces the blurred distribution, until the field converges. The smoothing operator used for N3 correction is a B-spline approximator. In the work by [214], this B-spline approximator was adapted to allow its use with larger field strengths, faster execution times, higher levels of frequency modulation of the bias field, and to be more robust to noise. Additionally, it utilises a multiresolution optimisation which iteratively corrects the bias field and estimates the residual bias field.

2.2.2 Skull-stripping

One of the first steps in the neuroimage processing pipeline is often the removal of extrameningeal tissues, such as eyes, fat and the dura from the whole head MR images. Since the succeeding image analysis steps in the pipeline use the skull-stripped images for further processing, their accuracy is influenced by the quality of the skull-stripping method. Potential over- or under-segmentation of brain tissue can lead to errors such as imprecise estimations of tissue volumes or its spatial distribution. The proposed methods can be classified in two main categories, edge based and template based. Edge based techniques aim at finding an edge between brain and non-brain structures. Edge based techniques are employed by the brain surface extractor

(BSE) [186] which uses an anisotropic diffusion filter, a Marr-Hildreth edge detector, morphological filter and region growing. The brain extraction tool (BET) [191] estimates a brain/non-brain intensity threshold from the image histogram and calculates a rough estimate of the centre of gravity and the radius of the brain. These estimates are used to initialise a spherical tessellated surface model. Each vertex of the surface iteratively undergoes an incremental movement. The movement is governed by an intensity term that finds a local threshold between brain and non-brain intensities which also takes the global thresholds into account. Freesurfer [67] uses a hybrid approach [200]. Similarly to BET, general parameters such as intensity thresholds for CSF and white matter, the centre of gravity and the brain radius are estimated. In addition, the global WM minimum location is determined, which is used as the main basin for the subsequent watershed algorithm. This results in a first estimate of the brain volume and is used to initialise a deformable surface algorithm. The active contours method iteratively deforms a template while using the brain volume from the watershed algorithm as a mask. The resulting surface is compared to a statistical atlas created from manually segmented images, which allows further refinement in case of segmentation errors. More recent algorithms also use convolutional neural networks [121], which can be trained and used on different modalities and pathologically altered head scans.

The second group of methods uses a deformable atlas with expert delineations which is registered to the target. This makes it more robust to intensity inhomogeneity and different acquisition protocols. A hybrid method that combines the discriminative attributes of a Random Forest classifier to detect the brain boundary and the generative attributes of a point distribution model is ROBEX [104]. Offline, the Random Forest classifier is trained on a set of features derived from the voxel intensities, after intensity inhomogeneity correction and intensity normalisation. The final classifier yields a probability volume, which indicates the probability of each voxel to be part of the brain boundary. A point distribution model is constructed from a set

of training images with corresponding landmarks. The result is a model of possible shapes which can be deformed according to the active shape model explained in Section 1.2.1.2. At runtime, a new target undergoes the same pre-processing steps such as intensity normalisation, feature calculation and voxel classification. Conceptually similar to single-atlas based segmentation, the point distribution model is fitted to the probability volume of the target with a non-linear deformation to refine the alignment. The final step involves the construction of the binary volume.

Depending on the registration approach used, MAS-based methods can be computationally more expensive and time-consuming. For example computational time is a limitation of Doshi *et al.'s* approach [73]. They used DRAMMS [150] for the nonlinear alignment of a selected study-specific template to the target because of its robustness to intensity inhomogeneities, noise, and outliers or missing anatomical regions. Eskildsen *et al.* presented a faster alternative [77], which is based on the nonlocal means patch-based approach and requires only linearly aligned atlases.

2.2.3 Tissue classification

In SPM spatial image normalisation (i.e. registration to a standard space template), tissue classification, and bias correction is combined in a unified model [15]. It is assumed that the distribution of voxel intensities in each set of predefined tissue classes is normal or a mixture thereof and can be described with its respective mean and variance (mixture of Gaussians). The spatial probability of GM, WM and CSF is provided by the ICBM MNI 152 atlas in a normalised stereotactic space and it is assumed that intensity inhomogeneity, i.e. a smooth varying field, is present. Due to the many unknown variables, an iterative algorithm with initial estimates, given by probability maps and a uniform bias field, was implemented. Each iteration starts by calculating the cluster parameters from the probability maps and the bias field. Then the probability maps and the bias field are updated until

convergence. In each iteration, the probability maps change slightly towards the distribution of the target image. In detail, the cluster parameters are calculated by computing the number of voxels that belong to each cluster, which further allows the calculation of the weighted mean of the image voxels and the variance for each cluster. From these cluster parameters, the prior proability images, the bias corrected field and the new label probability maps can be constructed following Bayes' rule. The smooth varying bias field is modulated as a linear combination of low frequency discrete cosine transform basis functions.

FSL's FAST tool [244] incorporates both a hidden Markov random field (HMRF) model and an expectation-maximisation (EM) algorithm. Over models fully specified by the histogram (finite mixture models), the MRF model has the advantage of taking spatial and, in turn, structural information into account. The spatial information is incorporated by a contextual constraint via a neighborhood system. The HMRF is an extension to this MRF and indicates that the states of the random field can not be directly observed. However, given a particular neighbourhood configuration and its parameters, a known conditional probability distribution can estimate them. The algorithm iterates through three EM steps, which include estimation of the class labels by MRF and maximum a posteriori (MAP), estimation of the bias field performed by MAP, and estimation of the tissue parameters performed by maximum-likelihood (ML). It is initialised with the mean and standard deviation for each class type, GM, WM and CSF, extracted from the histogram with Otsu's thresholding method [149]. Since MRF and HMRF are not the main focus of this manuscript, I would refer the reader to the above mentioned literature for more details.

Like FAST, ANTs' Atropos [22] is also based on a finite mixture model (FMM). However, a FMM assumes independence between voxels and does not take spatial and contextual considerations into account. Atropos uses prior probability models to incorporate this information and provides three options. Similar to FAST, a MRF can be used. Another way is to register

and warp labelled templates to the target to create a prior probability map. A third option in Atropos allows the user to label (sparse) points of the image, which provides an initialisation for the subsequent EM optimisation. In the E-step the label estimates for each voxel are updated based on the current estimates of the model parameters by computing the lower bound of the objective function. In the M-step the estimates of the model parameters (mean,variance) are updated by optimising this bound.

2.3 Image registration

Medical image registration has been an active area of research for over 20 years with the goal of estimating the spatial correspondence between anatomical structures in one image and the corresponding anatomical structures in a second image. The result is a transformation which indicates the correspondence between voxels across both images. This correspondence can be established by superposing landmarks or by matching voxel intensities. Landmark-based methods look for the same features in the images and try to align them with a transformation. Intensity-based methods try to increase a similarity measure. It is important for both approaches to define valid constraints that govern the process of finding the best images [85]. Although a one-to-one mapping between each of the voxels of two images would yield a perfect registration, in practice, it is very difficult to achieve because of intensity inhomogeneities, partial-volume effects, tissue motion, pathology, artefacts or simply because anatomy varies between subjects [65]. The evaluation of registration quality can be based on the use of labels [122] or features, automatically derived from the input images [193]. Both approaches are prone to errors because manually denoted labels might be inconsistent due to inter- and intra-subject variability and automatic features might not be correctly detected in complex structures with a high anatomical variability like the brain.

An image registration method based on intensities usually consists of three

main components:

- Similarity or error measurement
- Optimisation function
- Transformation model

In the following sections an outline of these key elements will be provided. For further details the reader may refer to [196].

2.3.1 Similarity measure

Similarity measures can be broadly categorised into three main groups: Feature-based, voxel-based and hybrid methods. Feature-based measures use geometric landmarks including points, lines or surfaces aiming at decreasing the (typically Euclidean) distance between corresponding features. These landmarks can be manually placed or automatically detected. Consequently, the registration accuracy largely depends on the ability of the feature extraction method to find corresponding landmarks in the images. Voxel-based measures are based on image intensity values with the goal to maximise the similarity of corresponding intensity patterns. Due to this dependency, we can distinguished between intra- and inter-modal registration. Common measures of similarity for images of the same modality include sum of squared differences (SSD), sum of absolute differences (SAD) and cross correlation (CC). Multi-modal image registration methods usually employ probability based methods, e.g. mutual information, which allow the comparison of images in terms of information theory (entropy) with the goal to maximise the amount of information one image yields on the other [55, 138, 230]. Another approach to inter-modal registration employs a patch-based method (see Section 4.8) to measure similarity between patches in a local neighbourhood of an image and detect preserved anatomical structures [101].

2.3.2 Transformation

The transformation is the mathematical model that describes the geometric distortions that are applied to an image. Based on the chosen model and its mathematical constraints, the transformation can gain properties such as inverse consistency, symmetry, topology preservation or diffeomorphism. Inverse consistent and symmetric methods ensure that the path along which an image is transformed to a second image is unbiased by the computation direction or space of either one of the images. Consequently, the path used when registering an image A to image B is the same path when registering image B to image A unbiased by the target domain. Another desirable property for some applications is the transformation model's ability to preserve topology. By creating a continuous and locally one-to-one mapping with a continuous inverse, folding of the grid over itself, resulting in potentially unnatural anatomical structures, can be prevented. This can also be achieved with diffeomorphisms [51, 211], which are transformation functions that are differentiable (smooth) and invertible with smooth inverses.

One linear model is the rigid-body model, which only allows rotations and translations (6 DOFs). It is a special case of the affine model, which additionally allows scaling and shearing (12 DOFs). These are commonly used to correct for global shape changes, but are too constrained to describe local shape changes.

In contrast, nonrigid or nonlinear registration methods have more DOFs and can be applied to model local tissue deformations, align anatomical structures that vary within a population or quantify change over time. Some of the most commonly used registration methods are based on physical models and interpolation theory [196]. The former group includes elastic body models, diffusion models (Demons approaches) and flows of diffeomorphisms, while free-form deformations are an example of the latter.

Elastic body models assume that the image has similar properties to an elastic solid material, which can be modelled with the Navier-Cauchy partial differential equation [26]. The equation is governed by internal and external forces. Based on the chosen similarity measure, the external force causes the deformation, while the internal force models the stress within the elastic material. Linear elastic registration methods cannot handle large deformations because the internal force of the equation increases proportionally with increasing external force making it only valid for small displacements. For large deformations, nonlinear elastic models were presented [153, 159].

Demons approaches model the transformation as a diffusion process [205]. The method considers the object boundaries in one image as semipermeable membranes, which allow the object of the second image to diffuse through them. Points on the membrane are called demons and decide whether a point of the moving object should be pushed inside or outside by iteratively computing each demon's force and updating the transformation based on all these individual forces. One way to estimate the forces is with the optical flow equation. Thirion's Demons approach has built the basis for a whole group of diffusion-based methods, most of which are based on the same iterative approach of estimating each demon's forces and updating the transformation based on these forces. The original model did not ensure diffeomorphic transformations which was later implemented by Vercauteren et al. [218].

Flows of diffeomorphisms estimate a displacement by integrating a velocity field over time with the Lagrange transport equation [51]. The velocity field can vary over time, which allows the esimation of large deformations as a composition of a series of small deformations [29]. Due to the integral of the velocity field along a path the deformation field is necessarily diffeomorphic which is why this framework is also known as large deformation diffeomorphic metric mapping (LDDMM). The shortest path in the space of diffeomorphisms is called the geodesic path which allows the comparison of

distances between points or images.

DARTEL (Diffeomorphic anatomical registration using exponentiated Lie algebra) [14] is based on diffeomorphisms and allows rapid computation of the deformations due to its use of a constant velocity field and a full multigrid method which recursively goes through the scales. The use of a stationary velocity field, where the whole movement of a point over a series of time steps is integrated into a single fixed velocity field, does not allow relating each point in the flow field to the corresponding point in the brain at each time step. Starting by calculating the first and second derivatives of the objective function for a variable velocity vector field, it is thereafter constrained to constant velocity. This constraint limits the achievable diffeomorphic configurations.

ANTs-SyN: Avants *et al.'s* symmetric image normalisation method (SyN) [21] extends the initially asymmetric LDDMM by guaranteeing that the geodesic mapping between two images is symmetric for every chosen similarity measure, not only for intensity differences. This is achieved by decomposing the diffeomorphism that deforms an image into a second image into two parts. Each of the images contributes equally to the whole deformation by constraining the sub-diffeomorphisms to be the same at half of their respective integration time, which is equivalent to half of the respective distances of the whole geodesic path. Local cross correlation was chosen as similarity measure, due to its robustness to intensity variations.

Free-form deformation (FFD) methods use a mesh of control points which can be manipulated with spline functions and consequently deform an object in the image. Different types of splines have been used where cubic B-splines [177], where adjusting the location of one control point impacts only a local neighbourhood of points, have become the most popular. FFDs have been extended by adding properties such as topology preservation and diffeomorphisms [176], symmmetry [148] and inverse consistency [79].

In a comparison by Klein *et al.* of 14 nonlinear registration methods [122] DARTEL and SyN were amongst the highest ranked methods in terms of overlap- and distance measures.

2.3.3 Optimisation

In most methods, the transformation is iteratively refined by estimating new transform parameter values to optimise the similarity measure of the images in the next iteration. This gradual improvement and subsequent similarity assessment is repeated until an optimum is found. The overall goal of the optimisation algorithm is to find the global optimum. One way to find an optimum is by using gradient-based techniques (gradient descent), which, however converge to a local optimum. A starting estimate close to the global optimum can be provided by reducing the capture range with a multi-scale method. By down-sampling, and then repeatedly up-sampling and registering the images, the transformation at each resolution level can be used as an initial estimate for the next registration step.

2.4 Anatomical atlases and label probability maps

Due to image artefacts such as intensity inhomogeneities, partial volume effects or similarities in intensity distributions of different anatomical structures, prior information is crucial for the automated segmentation of MR brain images. In atlas-based segmentation methods, this prior information is provided in form of a training dataset annotated manually by experts, which classifies it as a supervised learning method.

The first step in atlas-based segmentation requires the building of one or multiple anatomical brain atlases and their corresponding labels. It is crucial to clearly define what is considered as a brain atlas, as depending on the construction and application, scans from individual subjects and probabilistic

templates are often referred to as atlases in the literature.

2.4.1 Single-subject atlases

In earlier publications an atlas represented the annotation of anatomical structures and was called topological, single-subject or a deterministic atlas, with one of the most well-known examples in medicine being the Talairach atlas [201]. Talairach's main work aimed at describing and locating anatomical structures not only based on their shape but also based on their location relative to each other. In his coordinate system he used the anterior commissure to the posterior commissure for alignment and additionally introduced a parallel and orthogonal grid proportional to the skull size, which already provided the basis for the reconstruction of 3-dimensional volumes from 2-dimensional projections and is still widely used. With advances in imaging methodologies, more complex atlases have been developed. One representative example of a deterministic whole body atlas is the Visible Human Project. CT and MRI images were obtained in addition to cryosection images from whole human female and male cadavers, resulting in a complete digitial image dataset of the human body.

In a similar project, the Computerised Brain Atlas (CBA) was constructed from cryosections with the main goal to provide a brain template for normalisation with positron emission tomography or other imaging modes [94]. It contains 3-dimensional annotations of the brain surface, the ventricular system, the cortical gyri and sulci, as well the Brodmann cytoarchitectonic areas.

More recent atlases have been developed without the need for cryogenic images. Instead they have solely been constructed from non-invasive imaging modalities such as MR or CT scans [27, 117]. A high-resolution brain atlas was constructed by the McConell Brain Imaging Centre and used for the

BrainWeb database, which is a simulator for the creation of realistic MRI data volumes [58]. With the intention to create an average atlas with high SNR and structure definition, the Colin 27 brain atlas [102] was constructed from 27 linearly aligned scans of the same subject. However, after bringing it into stereotaxic space, it has also been frequently used as a template for alignment.

2.4.2 Population atlases

Due to the large anatomical inter-subject variability, there is no single brain anatomy scan capable of representing a whole population. Probabilistic atlases, constructed from a set of images, have emerged as the tool of choice in the representation, analysis and interpretation of population-based imaging studies.

Population atlases provide a common 3D (stereotaxic) space for normalisation and a reference for alignment. By mapping images from a cohort into the same space, population atlases provide a probabilistic map of the spatial location of structures of interest and an estimate of their shape (intra-population variability), allow the estimation of morphological differences between distinct populations (inter-population differences) and provide support in the segmentation of new target images.

The use of probabilistic atlases has multiple advantages over individual-subject atlases:

- Since it represents the average shape and appearance of the population, it provides a space for normalisation for both the individual images of the population and the targets. For the individual images, it requires the least amount of deformation to deform to their average. Consequently, for targets with similar morphological properties to the population, the quality of the registration can be increased, resulting in more precise deformation fields.

- Probabilistic atlases can be constructed for different populations, such as cohorts of different age, gender or pathology [241]. For an individual target image, the most suitable probabilistic atlas, which requires the least amount of deformation, can be selected for spatial normalisation [162, 169, 239]. On population level, the probabilistic atlases can be compared to find morphological differences.

- Three-dimensional spatial atlases can incorporate information from additional target images later on and can be extended into the time domain, which makes it possible to compare diseases at different points in time or investigate developmental disorders [90, 156, 184].

The main goals can be defined as finding a realistic representation that captures the structural and functional variability of a cohort individual scans and allowing the quantitative estimation of accuracy and errors [144].

One of the biggest attempts to create an atlas was made by the International consortium for brain mapping (ICBM) in a worldwide collaboration of imaging centres. Demographic, clinical, behavioural and imaging data of a total of 7000 subjects were collected, including genetic information from approximately 80% of the subjects. The processing steps for the construction of probabilistic brain atlases from 152 and 452 subjects included 3-dimensional intensity non-uniformity correction, and intensity normalisation. In order to create an average atlas that is spatially as well as intensity unbiased by a single subject, they were constructed from the average position, orientation, scale, and shear from all individual subjects [144, 145, 146].

The Laboratory of NeuroImaging (LONI), also part of the ICBM, constructed a probabilistic brain atlas from 40 manually delineated MRI scans of healthy volunteers. The delineation was performed for 50 cortical structures, 4 subcortical structures, the brainstem, and the cerebellum. These label maps were brought into a common space, which allowed the calculation of proba-

bility density functions for each structure [185].

Similarly, the Internet Brain Segmentation Repository (IBSR) [1] comprises 18 T1-weighted MR images of the whole brain and their corresponding manual segmentations, with 32 labels per map and the Non-rigid Image Registration Evaluation Project (NIREP) [49] provides a set of 16 topological atlases of healthy subjects with 32 annotated grey matter regions each for the evaluation of software tools such as non-rigid image registration algorithms.

In order to investigate differences in regional and total brain volumes between preterm and term-born infants, a set of manually segmented topological atlases called ALBERT was constructed [92]. The 15 infant scans were delineated into 50 regions each.

While most of the previous projects focused on the construction of atlases from healthy subjects, there have also been datasets for different pathological cohorts. A representative example includes topological atlases of patients with Alzheimers disease from the Open Access Series of Imaging Studies (OASIS) [141]. It contains a total of 416 subjects of which about 100 were diagnosed with Alzheimers disease (AD). A subset of images from healthy controls was used for the MICCAI 2012 [125] and MICCAI 2013 [17] challenges on MAS. An even larger repository of clinical and imaging data is provided by the Alzheimers Disease Neuroimaging Inititative (ADNI) [109], containing data from over 1880 subjects. However, manual segmentations are only available for a subset.

In addition to cross-sectional datasets, both repositories also provide longitudinal data from a part of their cohorts, including nondemented and demented

[1]The MR brain data sets and their manual segmentations were provided by the Center for Morphometric Analysis at Massachusetts General Hospital and are available at http://www.cma.mgh.harvard.edu/ibsr/.

subjects with scans from at least two visits [140]. By adding time to the spatial 3D domain, 4D atlases can be created. The time domain allows to capture growth or disease progression and prediction making. The cross-sectional and longitudinal information allows the classification and comparison of healthy subjects to Mild cognitive impairment, cognitive impairment and AD. In general, population atlases also allow the classification by other attributes such as age or sex, leading to the construction of atlases based on group-specific criteria, which make it flexible enough to incorporate new images into the atlas later on. Furthermore it facilitates the comparison of parameters such as volume and shape of anatomical structures on population level.

2.4.3 Unbiased probabilistic atlas construction

Since there is no single subject brain scan capable of representing a whole population and its large anatomical inter-subject variability, probabilistic atlases, constructed from a set of images, have emerged as the tool of choice in the representation, analysis and interpretation of population-based imaging studies. The optimal unbiased probabilistic atlas most representative of the dataset is the one that requires the minimum amount of deformation to all individual atlas images of the dataset. In the simplest case, a probabilistic atlas can be computed as the mean image of a collection of images, thereby representing the average anatomy of the population. Within the context of this definition, a population atlas refers to a 3D average of cross-sectional data without any notion of time or age of the subjects.

In early methods, the anatomy of a single subject was used as a template for normalisation. All other population images were then brought into this same space. A limitation of this approach is that the choice of template introduces a bias towards the shape of its anatomy. One solution towards the use of unbiased brain atlases for determining differences in anatomical patterns was proposed by Thompson and Toga [209]. A target image is registered to all existing atlases and the resulting deformation fields are used to

determine the probability distribution of corresponding points. This allows the calculation of the likelihood of the targets region to be in a similar spatial location to the corresponding atlas brains. Due to the warping of the target to all atlas images, all of them contribute equally. In order to avoid the registrations between each atlas and a new target, an unbiased atlas can be calculated by simultaneously registering all of the subjects to their average coordinate system [33]. This makes the selection of a reference subject unnecessary. In this approach the computation of the arithmetic average of small displacement fields (vector fields) is well defined. This is not the case in the large deformation setting where we operate in the high dimensional group of diffeomorphisms. Based on the theory of large deformation diffeomorphisms, which allows the estimation of the displacement by integrating a velocity field, a distance or similarity measure is defined and the problem can be stated as finding the image and associated coordinate system that requires the least amount of deformation to align with every subject of the population [113]. The solution for this minimisation problem can be estimated with a greedy fluid algorithm by iteratively updating the transformation for each image and updating the template as the voxel-wise arithmetic mean. The final deformation can then be composed of the transformations gained at each iteration. One drawback of this method is that the average arithmetic mean of different tissues that are not in correspondence might lead to biologically wrong structures.

In order to solve this problem, SPM [14, 16] first requires approximate alignment of the images to pre-defined tissue probability maps. Once in the same space, tissue probability maps for GM, WM and CSF can be constructed for each of the individual images. These individual subject maps are iteratively registered to their average followed by the construction of a new average. The initial template is calculated as the intensity average of the GM and WM maps. The similarity measure for the registration is based on the SSD, which considers the difference between the likelihood of the individual GM probability maps and the GM mean, between the likelihood of

the individual WM maps and the WM mean and the likelihood of the rest (1-prob(WM)-prob(GM)). Due to this procedure, the template construction and registration is not directly based on the image intensity values but rather on the tissue probability maps. Compared to the first mean, which is very smooth and blurry, the template becomes crisper with every iteration. The registration of each image provides a diffeomorphic deformation field to the average. On the one hand, the initial alignment to the pre-defined tissue probability maps and the concurrent use of the maps does not introduce false structures when averaging them. On the other hand, this initial alignment restricts the anatomical space, i.e. the space used for normalisation is not the average of the individual images.

In the advanced normalisation tools (ANTS) package, a method for the symmetric-group-wise normalisation (SyGN) independent from an individuals space and unbiased by an individuals shape was proposed [23]. The template creation facilitates the use of the previously described SyN algorithm for registration. Diffeomorphisms with symmetric properties are computed for the alignment of the individual images and the algorithm does not require an initial template estimate. Thereby, both shape and appearance of the constructed template are unbiased by the individual images. The goal is stated as finding the template and transformations that increase image similarity and reduce the path length of the diffeomorphisms as stated by an energy function. The diffeomorphisms that map a template to the individual images and the template shape as given by another diffeomorphism are initialised as identity. Then each part of the energy function is optimised while the rest are kept constant. First the diffeomorphisms between each image and the fixed template are re-estimated. Then the template appearance is updated with fixed shape and diffeomorphisms. And finally the template shape is updated. The template used in the first iteration is the affinely registered average appearance image [194]. In more detail in one iteration the set of transformations between each image and template is computed. Then the template appearance is iteratively calculated by deforming each of

the images with their corresponding inverse transformation. The gradients of the similarity term are calculated for each deformed image and averaged, which is followed by an update of the template. The shape is updated by averaging the diffeomorphisms between the template and each image. The new diffeomorphism can be applied to the template deforming its shape more towards the new average.

2.4.4 Creation of label maps

Label maps are usually acquired by means of accurate manual delineation which is a very time-consuming task but has to be done only once for each image and can then be re-used without further manual interaction.

The preprocessing steps for the creation of the LONI LPBA40 population atlas [185] require the alignment of the individual images to delineation space for which the ICBM MNI-305 space [78] was chosen. The registrations included the identification of ten landmarks in each image and the MNI brain atlas followed by rigid-body translations and rotations. The transformations with six degrees of freedom were applied to the corresponding images to bring them into MNI space. Bias fields were corrected with the N3 method and the brain was extracted with FSL BET. In the labelling process an expert assigns a pre-defined integer value to every voxel of pre-defined ROIs. Each value indicates the presence of a particular neuroanatomical structure at the specific location. In order to increase the overall accuracy and keep inter- and intra-observer variability to a minimum, trained experts follow a set of protocols that specify the structure to be delineated, a single plane of section (coronal, sagittal, axial) in which the structure should be labelled, and details about grey matter, white matter and sulci that should be included or excluded. These instructions are based on anatomical landmarks and provide visual samples with 3D cross-sectional views. Additionally, each of the segmentation steps with starting and end points are given. In the LPBA40 the grey matter was included in cortical structures. Pairs of experts were

assigned to each anatomical structure and performed the delineations independently. The calculated volumes of both experts were compared for each structure using the corresponding Intraclass Correlation Coefficient (ICC) given as $ICC = \frac{MSB-MSW}{MSB+MSW}$ with MSB being the mean squares between the volume measurements of the structure and MSW the mean squares of the volume measurements within each structure. In order to become a certified rater, the ICC scores had to exceed a certain threshold or rater pairs were retrained until a reliable result was achieved. In addition to the ICC, the Jaccard similarity index [108] was calculated. In contrast to the ICC, which assesses the interrater variation of the volumes, the Jaccard index provides a measure of overlap of the volumes (Section 2.5). Once a reliable ICC and overlap was achieved, the same structure was segmented in the remaining subjects. For voxels in overlapping or missing areas between structures, pairs of raters decided on and assigned one label.

Similar to the LONI LPBA40, the Harmonized hippocampal Protocol (HarP) [35], which has been applied to a subset of the ADNI dataset consisting of 1.5T and 3T images, includes a series of pre-processing steps to align the subjects in the same space. Each hippocampus was segmented by five different raters and ICC and Jaccard overlap scores were calculated from their segmentation results of an independent sample of 20 subjects. The segmentation protocol comprises a set of guidelines for the delineation of the outer contour followed by filling the area encapsulated by it. This is done in the original space of the images with sub-voxel accuracy. In order to refine the contour, the resolution of the initial grid dimensions is increased by a factor of 10. These high-resolution images can then be resized into the initial low-resolution space. Due to the use of interpolation methods the resulting segmentation consists of probabilistic values between 0 and 1. For the final segmentation the images were binarised with voxel values of less than 50% being discarded.

The segmentation protocol of the NIREP dataset [49] does not provide details about the raters or training, but refers to the literature for the delin-

eation of neuroanatomy and describing its variation [7, 8]. The segmentation was performed in 2D, which makes for smooth delineations in the segmentation plane but rough edges viewed from other directions. In order to avoid boundary errors, the initial segmentation, which also contained white matter, was restricted to grey matter alone leading to ROIs with smooth boundaries on the surface and between grey and white matter.

2.5 Tools, datasets and evaluation strategies

For the implementation and validation of the framework presented in this document the following tools and datasets were used. Images were re-oriented to match the orientations of MNI space using FSLs reorient function. Skull-stripping was performed with BET, ROBEX or Freesurfer. Based on visual evaluation, overall Freesurfer showed the best performance and was subsequently used for all datasets. Images were affinely registered to the MNI-ICBM 152 brain atlas to bring them into a common space with uniform dimensions and voxel size. The nonrigid registration and population template construction was performed with SPMs DARTEL or ANTs SyN. Due to the higher computational complexity of SyN, for most of the evaluation DARTEL was used as specified in the following chapters. Extensive evaluation was conducted on the LONI-LPBA 40, IBSR, NIREP, ADNI with HarP protocol and MICCAI 2012 and MICCAI 2013 datasets (Table 2.1).
All datasets provided individual atlases consisting of an intensity image and a label map each. Each of the datasets was randomly divided into atlases and test images for cross-validation. From each atlas subset a population atlas was created. The corresponding test images were sequentially nonlinearly registered to the population atlas. This is in large contrast to other recent methods [31], which used atlases and test images for the construction of the population atlas. However, considering that a bias towards the test images is introduced in its appearance and the accuracy of the corresponding deformation fields, the population atlas would have to be reconstructed

Dataset name	#of subjects	Image dimensions	Voxel dimensions	#of regions	Age range	#of males	Scanner	Scan protocol	TR [ms]	TE [ms]
LONI-LPBA40	40	256x256x124	0.86x0.86x1.5 (n=38) 0.78x0.78x1.5 (n=2)	54	19-39	20	GE (1.5T)	SPGR	10.0-12.5	4.22-4.5
ADNI-HarP	135	n/a	slice thick.=1.2	2	60-89	70	Siemens, Philips, GE, 1.5T & 3T	MPRAGE	n/a	n/a
IBSR	18	256x128x256	0.84x1.5x0.84 (n=4) 0.94x1.5x0.94 (n=8) 1.0x1.5x1.0 (n=6)	32	7-71	14	GE Signa (1.5T)	SPGR	40	5
NIREP	16	256x300x256	1.0x1.0x1.0	33	24-48	n/a	GE Signa (1.5T)	SPGR	24	7
MICCAI 2012	35	256x256x128	1.0x1.0x1.25	113	18-90	10	n/a	MP-RAGE	n/a	n/a
MICCAI 2013	35	256x256x287	1.0x1.0x1.0	14	15-96	10	Siemens Vision (1.5T)	MP-RAGE	9.7	4

Table 2.1: Characteristics of the used datasets.

for every new target image which is very time-consuming. Our evaluation strategy was guided by a real life scenario where test images would not be present at the time of population atlas creation. Consequently, it was constructed only once from the atlas images and re-used for every new target thereafter. The evaluation accuracy of individual label overlaps is also influenced by the way multiple adjacent labels which might be overlapping in the final segmentation results are handled. One way would be to consider each label as a separate entity with its label probability values. Another way is to combine individual labels into a single segmentation map. In the latter case overlapping areas would have to be assigned to one of the competing labels based on their label probability value. In our evaluation the second approach was chosen to provide a single label map consistent with the atlas label maps used as input.

Due to the reliance of the MAS concept on image registration, both the evaluation of nonrigid registration methods and the quantification of segmentation accuracy are very closely related. In the literature, registration methods are commonly assessed with surrogate measures such as anatomical label overlap, tissue overlap, image similarity, and inverse consistency [195]. However, Rohlfing showed that only overlap measurements of small labelled ROIs could assess registration quality accurately [167]. Consequently, in our experiments the overlap of a label annotated by means of expert manual segmentation, which is considered as the gold standard, and the label assigned by the segmentation method under assessment was used for performance evaluation. Usually the region overlap is given either by the Jaccard similarity (JS) [108] or the Dice similarity coefficient (DSC) [71], both shown in equation 2.1 and related by $DSC = \frac{2*JS}{JS+1}$ [66]. The JS of two overlapping regions A and B is defined as the ratio of the size of the intersecting volume and the size of the union of the volumes. The DSC can be written as the ratio of the intersecting volume and the mean of the volumes. Applied to medical images the respective volumes are given by the enclosed number of voxels $N()$.

$$JS(A,B) = \frac{N(A \cap B)}{N(A \cup B)} \qquad DSC(A,B) = \frac{2 * N(A \cap B)}{N(A) + N(B))} \tag{2.1}$$

2.6 Our approach to template building

The first steps of our proposed method include pre-processing such as skull-stripping, affine normalisation to a standardised space and tissue classification of the MR brain images. The resulting tissue maps from the atlases are used to create an average population template.

Let us consider an a priori set of 3-D atlases $A = \{A_j = (I_j, \{L_j^r\}_{r=1...R})\}_{j=1...N}$, where each atlas consists of an intensity image I_j and R label maps $\{L_j^r\}_{r=1...R}$, with $L_j^r(x) = 1$ (maximum probability) if voxel x belongs to label r in atlas A_j and 0 otherwise (Fig. 2.1).

2.6.1 Pre-processing

We also ensure that $\forall x, \sum_{r=1...R} L_j^r(x) \leq 1$, i.e. the label maps do not overlap.

Next, we skull-strip all atlas images, $\{I_j\}_j$. BET, ROBEX and Freesurfer were tested with default parameters. The skull-stripped images are linearly registered with 12 degrees of freedom to a common space, that of the MNI152 MR brain atlas, using FSL FLIRT4.1. We obtain a set of affinely registered images, $\{\widetilde{I_j}\}_j$, and the corresponding affine transformations. We then estimate tissue maps, $\{\widetilde{G_j}\}_j$, $\{\widetilde{W_j}\}_j$, and $\{\widetilde{C_j}\}_j$, with SPM's New Segment. Finally, we apply the affine transformations to the individual label maps to also bring them into the same MNI152 coordinate system: $\{\widetilde{L_j^r}\}_{r,j}$.

2.6.2 Estimating a population template

We constructed average population templates with ANTs' SyN template building method and SPM's groupwise registration algorithm DARTEL, both highly ranked in evaluation studies [122]. DARTEL iteratively creates in-

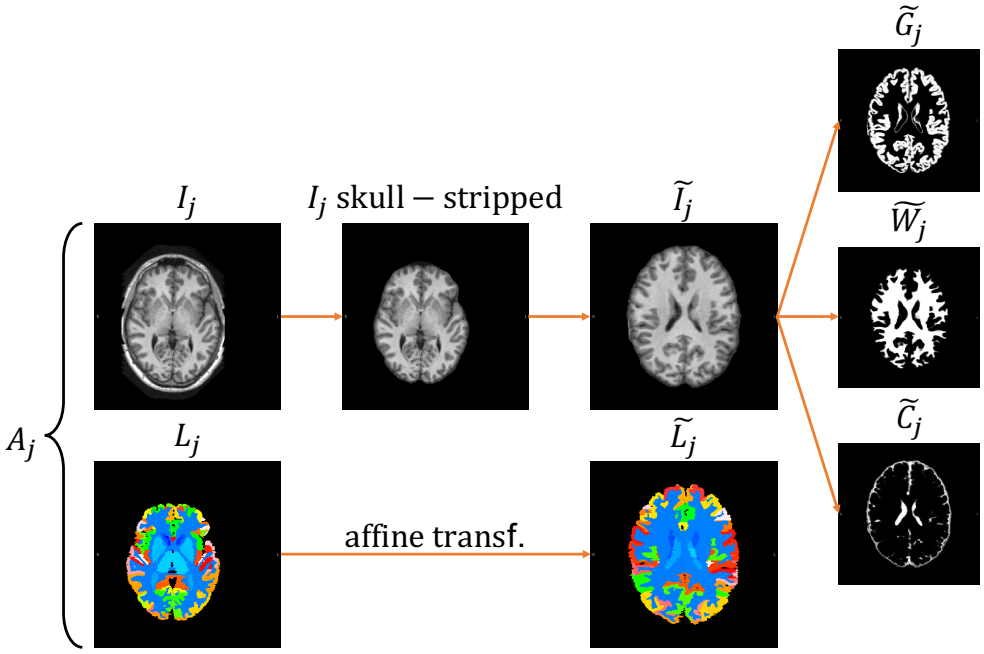


Figure 2.1: The atlas image in original space is skull-stripped, linearly transformed into MNI space and segmented into tissue maps. The same affine transformation is applied to the corresponding atlas-label map.

creasingly sharper population templates $\overline{I}$, $\overline{G}$, $\overline{W}$ and $\overline{C}$ from the set of affinely registered atlas intensity images and tissue maps. As a byproduct we also get a set of deformation fields $D = \{D_{\widetilde{I}_j \to \overline{I}}\}_{j=1\ldots N}$ that precisely map each image and its tissue map to the corresponding population template. Note that DARTEL creates all three population templates concurrently, using the tissue maps to improve the accuracy of the process. Consequently, we get the same deformation fields between the atlas images and the image template as between the atlas tissue maps and their respective population templates. SyN does not explicitly require tissue classified maps. Consequently only one template which contains all tissue classes is constructed. Similarly to DARTEL the output consists of the deformation fields that map each atlas intensity image to the population template. The DARTEL image template with all tissue classes can be created as the average of the warped atlas images, to allow for easier comparison to SyN's templates. The warped

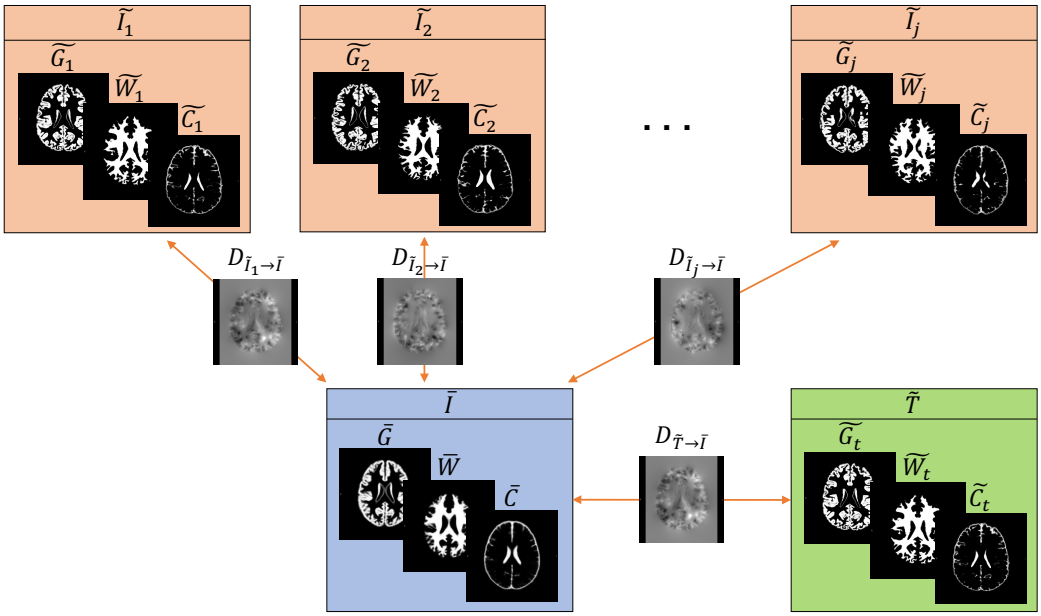


Figure 2.2: Offline population templates for GM, WM and CSF are concurrently constructed from the corresponding atlas label maps. At runtime the target is registered to these population templates. The created deformation fields map the template to each individual atlas and the target.

atlas images are obtained by applying each deformation field to the corresponding atlas image.

2.6.3 Transferring the label maps to the target

At runtime the same pre-processing steps as outlined in Section 2.6.1 are applied to the target image T to obtain an affinely registered target $\widetilde{T}$ in the space of the MNI152 MR brain atlas, and tissue maps $\widetilde{G}_t$, $\widetilde{W}_t$ and $\widetilde{C}_t$. These tissue maps are nonlinearly aligned to the population templates $\overline{I}$, $\overline{G}$, $\overline{W}$ and $\overline{C}$ by estimating the deformation field $D_{\widetilde{T} \to \overline{I}}$. The final transformation $D_{\widetilde{I}_j \to T}$ between each atlas $\widetilde{I}_j$ and the target T can be approximated by composing $D_{\widetilde{T} \to \overline{I}} \circ D^{-1}_{\widetilde{I}_j \to \overline{I}}$. The corresponding atlas label maps $\{L^r_j\}_{r=1...R}$ are transferred to the target by applying to them this composite deformation field.

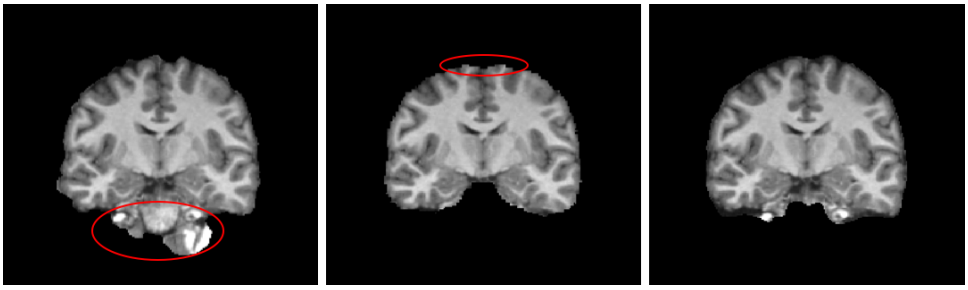


Figure 2.3: The same coronal slice of a subject's brain scan skull-stripped with FSL-BET (left), ROBEX (middle) and Freesurfer (right). While BET misses parts of the brain stem, ROBEX removes too much of the GM. Freesurfer showed the best trade-off with smooth borders.

2.7 Experiments

This approach was applied to the MICCAI 2012 dataset (Section 2.4.2) consisting of 15 atlas and 20 target images. The pre-processing steps, including skull-stripping and affine registration to MNI space, were performed for all images.

The visual comparison of the results of three skull-stripping methods (Fig. 2.3) showed large under-segmented but also over-segmented regions by FSL-BET. Here we define over-segmentation as the removal of too much tissue that should have not been removed and, conversely, under-segmentation as the removal of too little tissue leaving parts that should have been removed. Under-segmentation was mainly observed around the brainstem while some scans were over-segmented where parts of gyri were removed. ROBEX performed consistently well in removing non-brain tissue, but we noticed slight over-segmentation. Small parts of the GM and in some scans WM were removed. Freesurfer provided the best specificity at high sensitivity without removing GM tissue.

The DARTEL template creation process iterates through two main steps. Firstly, it refines the population template and secondly, it refines the defor-

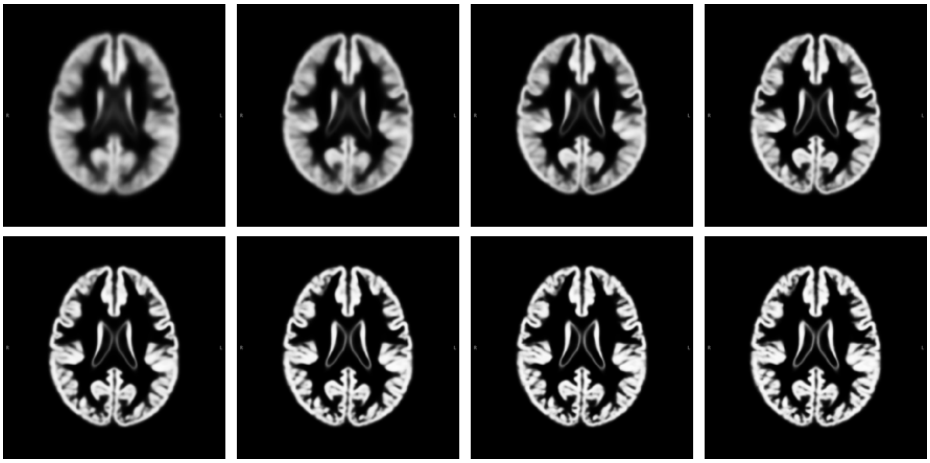


Figure 2.4: The same axial GM slice of the DARTEL template at different stages of the algorithm. With each iteration (from left to right and top to bottom) the population template, constructed from the individual atlases, becomes crisper and clearer.

mation fields. Initially it starts with a coarse average of all atlases, which becomes clearer and crisper with every iteration (Fig. 2.4).

A visual comparison of the final DARTEL template, created by averaging the warped atlas images, and the template, created by SyN, showed similar appearance (Fig. 2.5). SyN provided crisper borders with more details compared to the slightly blurrier template by DARTEL. The deformation fields between each atlas and the template, estimated with SyN appeared smoother compared to the deformation fields estimated with DARTEL (Fig. 2.6). However, in our runtime comparison DARTEL showed better performance than SyN on the same desktop PC.

2.8 Discussion

Pre-processing tools for intensity inhomogeneity correction, and brain tissue segmentation and classification, are often used in the first steps of the MAS pipeline. Consequently, their performance affects the quality of the final segmentation. It can have an impact on offline learning, image reg-

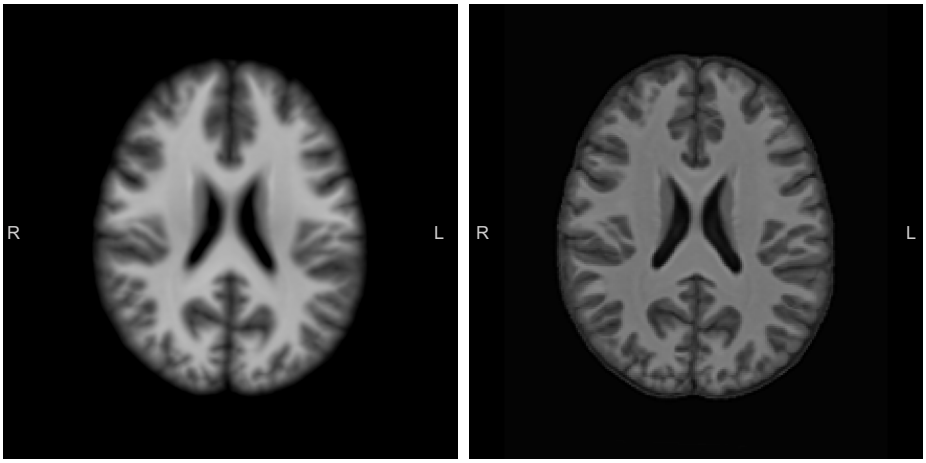


Figure 2.5: The same axial slice through the population templates created with DARTEL (left) and SyN (right).

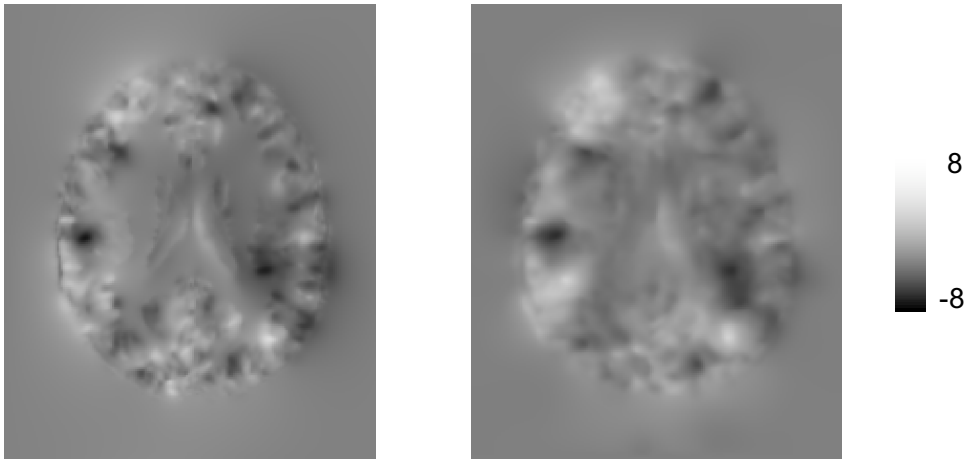


Figure 2.6: The same axial slice through the deformation fields in one direction, mapping a subject to the population template created by DARTEL (left) and ANTs (right). It is notable that the deformation field created with DARTEL is much crisper.

istration, atlas selection and fusion (Fig. 1.6), all of which require similarity measurement between images or features derived thereof. We tested three different skull-stripping methods, FSL-BET, ROBEX and Freesurfer, and found differences in the quality of the estimated brain masks. FSL-BET showed under-segmented brain regions with default parameters and,

even with manual parameter adjustments of the intensity threshold options, which allow estimation of a smaller or larger brain outline for the whole or parts of the brain, no ideal solution was achieved. ROBEX was the most user-friendly tool and, considering that it does not require parameter input, showed high precision in the region of the brainstem and in removing non-brain tissue. However, we noticed slight over-segmentation of cortical regions, which might introduce errors into the concurrent tissue classification and template construction. Freesurfer allows adjustment of multiple parameters for its combined watershed- and template-based approach and showed overall satisfactory results with its default configuration. Although the region around the brainstem was slightly under-segmented compared to the results by ROBEX, the cortical segmentation appeared more precise in our experiments. One of the challenges of skull-stripping is to find a robust method. Often the quality of the results is influenced by image contrast, artefacts, partial volume effects, anatomical variability and voxels with the same intensity values in brain and non-brain tissue. Consequently, the performance of skull-stripping methods might vary between different datasets or images, which makes a direct comparison and ranking challenging.

One crucial task in the MAS pipeline is image registration. Nonlinear registrations between the atlases and the target cause the main computational burden at runtime, which represents a disadvantage of many MAS approaches. Based on the degrees of freedom and the underlying approach, registration methods show varying registration accuracy and computational efficiency. In general, the direct estimation of a high-dimensional dense deformation field is computationally expensive, but can provide high registration accuracy of local geometric differences. MAS approaches using only affine registration methods with a small number of degrees of freedom, such as patch-based techniques, can align images much faster but show inferior segmentation accuracy compared to nonlinear patch-based approaches. In patch-based techniques the main computational burden is shifted towards the label fusion, where

the method iterates through voxels to assign a label to them. Other ways to improve computational efficiency include the use of better hardware and its parallel computing ability, which allows the computation of some tasks simultaneously. However, these resources are not commonly available to end-users yet and do not provide a solution to every task. Due to its computational efficiency, the use of a template for normalisation of atlas and target images has been of particular interest for our approach. Pre-computed population atlases are publicly available or can be constructed offline from a set of individual atlases. The use of the latter for MAS segmentation has shown to provide superior results, since it is can be constructed from images similar to the target. But current template-based methods still do not reach state-of-the-art segmentation results compared to direct registrations between each atlas and the target. We tested two of the most commonly used registration and template construction methods, DARTEL and ANTs. The comparison of the resulting templates showed crisper borders and more detail in ANTs' template compared to DARTEL's. Conversely, the corresponding deformation fields created with ANTs were smoother compared to DARTEL's, which could be related to the parameter configuration we used for DARTEL. This is further supported by Klein *et al.'s* study [122], where the accuracy of multiple methods was compared and both ANTs and DARTEL were highly ranked. We chose DARTEL over ANTs for template creation and concurrent registration steps because it showed faster performance in our tests. Given the importance of image registration in MAS, Klein *et al.* and other authors [65, 122] also noted that not every brain can be one-to-one mapped to any other brain and, consequently, image correspondence should not be mistaken for anatomic correspondence. They further concluded that registration accuracy can be improved by using similar templates as intermediary registration targets, but finding a template which provides high correspondence in every ROI might be difficult. Another problem is the measurement of registration quality. Rohlfing tested several commonly used accuracy measurement methods including the label overlap between the transformed and target la-

bels [167]. On a macroscopic level with large labels these measurements were not able to reflect the the degree of correspondence a registration method achieved. Rohlfing concluded that only the use of localised anatomical regions for overlap measurement can distinguish between the quality of registrations. This would require a dense set of landmarks, which is usually not available, or a set of small labels, delineated with high precision. The precise delineation of ROIs is also crucial for MAS, due to it its heavy reliance on "expert" segmentations. These labels are commonly acquired through means of time-consuming manual delineation. Although it is considered as the "gold standard" and detailed protocols for annotation are usually provided, they often suffer from inter- and intra-operator variability. In [105], Iglesias and Sabuncu pointed towards alternative methods to reduce the manual burden. Only those images with the largest potential to accurately segment a target could be manually delineated by experts. Conversely, Maier-Hein *et al.* [139] and Ganz *et al.* [86] showed that the delineation can also be performed by a larger number of "non-experts", which has shown to produce high-quality annotations similar to expert results. They used a crowd-based approach for feature annotation in endoscopic images. Similar results were presented by Bogovic *et al.* [36], where a hierarchical cerebellum parcellation protocol for non-experts was tested. However, the question remains whether this is applicable to more complex anatomical brain structures and scans of varying quality. Since they used high-quality research scans, it is questionable if a same level of accuracy can be reached in the presence of artefacts.

In conclusion, results from pre-processing methods should be carefully examined, since they can impact the accuracy of concurrent MAS steps. The use of a population template can improve computational efficiency at runtime and, when constructed from a set of atlases from the same cohort as the target, also improve quality of the target-template registration and segmentation accuracy. However, segmentation accuracy of traditional template-based methods are, in general, inferior to methods that use direct registrations. One

reason is that the target could still be very different to the population template in some ROIs, even when the template was constructed from the same population as the target [122]. The quality of the registration between target and population template can be improved by constructing an intermediary target-specific template (TST) at runtime.

Chapter 3

Target-specific template

3.1 Introduction

In the previous chapter the use of an average population atlas as a reference space for normalisation was presented. By using a reference space only one time-consuming nonlinear registration must be performed at runtime between the target and the average population template. The atlas images and the target are linked by composing their corresponding deformation fields allowing the propagation of labels from selected atlases to the space of the target where they can be directly fused [221]. However, the construction of the average popoulation template is performed offline, without the knowledge of the target. Consequently, if the set of atlases and, in turn, the population template are morphologically dissimilar to the target, the registration between the two images might be imprecise [168, 239].

With the similar motivation to reduce the number of registrations online but also to turn a potentially large, difficult registration task into easier, more accurate tasks, the construction of a graph or tree-like structure offline has been proposed [46, 88, 96, 110, 111, 118, 147, 202, 225, 234]. Each node represents an atlas image with deformation fields providing the links between them. The nodes are organised based on their similarity to each other and all deformation fields in the tree are pre-computed offline. A new target can be

efficiently added to the most similar atlas in the structure. The information can then be propagated from any node in the structure to the target by composing the deformations that link the corresponding nodes along the path. Each node along the path between the target- and an atlas node can be seen as an intermediate template to refine a potentially large deformation. One representative example is Wolz *et al.'s* [234] atlas propagation framework, where a manifold is learned from all atlases and unlabelled target images. Only similar atlases within a specified neighbourhood around each target were used for multi-atlas segmentation. A similar approach with the goal to reduce the risk of misregistration by refining large deformations into a sequence of smaller more accurate registrations via similar intermediate images was proposed by Gao et al. [88]. Each edge between nodes represents a weight defined by the pairwise image similarity. In contrast to Wolz's method, the information from all atlases was propagated along the optimal geodesic paths to the target images and a patch-based label fusion method was used to produce the final segmentation. A more generic model for the propagation of voxel-wise annotations between images in a spatially-variant graph structure was presented by Cardoso et al. [46]. A similar concept was used by Tang et al. [202] to improve registration accuracy between images. They used principal component analysis (PCA) on a training set of deformation fields, acquired by registering individual images to a selected template, to capture the statistical variation of the deformations. Each of the training deformation fields as well as the target deformation field, acquired by registering a target image to the same template, could be characterised by a small number of parameters in a low dimensional space. The most similar deformation fields from the training set were determined by comparison in this low dimensional space, which allowed the approximation of the target deformation field and construction of an intermediate template, further used for refinement. One criticism of this concept was identified in a study by Sjoeberg *et al.*'s [188] where the impact of intermediate templates for linking the target and the remaining atlases on the resulting segmentation quality was investigated for

a varying number of composed deformation fields and compared to a single direct registration. Images from a set of atlases with abnormal anatomy were randomly selected to serve as intermediate templates. It was shown that the composition of deformation fields in a multi-atlas framework can lead to a decrease in segmentation accuracy almost linear to the number of links compared to direct registration. Although results could be improved by ranking intermediate templates based on their similarity to allow more accurate registration, direct registration was superior. In the study, atlases of diseased patients with head and neck tumors were used, which makes the registration even more challenging. However, the use of composed deformation fields reduced computation time by approximately 2/3. Overall, they concluded that improved computational efficiency outweighs the differences in segmentation quality, making it a feasible method in a clinical setting.

An alternative method that aims to overcome the criticisms of using a population template or a graph-structure was presented by Commowick, Warfield and Malandain [59, 61, 161]. An average population template is constructed as a reference space from the atlases offline. At runtime the target is registered to this reference space. The locally most similar atlases to the target are identified by comparing their respective deformation fields under the hypothesis that more similar images should have similar deformations. A subset of these atlases is selected for the construction of a TST by iteratively registering the atlases to their average, constructing a new average template and applying the inverse of their average transformation to this template. This strategy represents the local extension to the average template construction method proposed by Guimond [95]. The target can then be registered to the TST. Assembling a TST from similar atlas ROIs is expected to yield a registration of high quality to the target since the locally most similar atlases to the target are selected, rather than the whole atlas images with no guarantee that those would be the most similar for each ROI. At runtime it requires only two nonlinear registrations, one between the target and the average population template and one between the target and the

TST. The atlas labels can be propagated to the target by applying the composed deformation fields between the corresponding atlas and the TST and between the TST and the target. The advantage of using a TST has also been shown for atlas building in the presence of artefacts [151] and for multi-organ segmentation [235]. Shi et al. presented a similar strategy for constructing neonatal brain atlases at different stages of development [187]. Individual atlas images are spatially normalised by creating a population template from them. This population template is parcellated into smaller ROIs with a watershed algorithm. The atlases are clustered into sub-populations for each ROI, identifying groups of similar shapes with one atlas in each group being determined as a representative exemplar for the group. For each ROI and for each group an average template is constructed. A given target is normalised to the initial whole brain average population template and for each ROI the most similar exemplar and consequently, sub-population is determined based on MI similarity. The TST is then assembled from the corresponding regional average templates. Due to the use of regional average templates, constructed from similar sub-populations, the final TST is more similar to the target than an average whole-brain template.

The comparison of images plays a crucial role in finding those atlases with the highest chance of producing a similar TST to the target and an accurate segmentation. This comparison can be performed with different metrics for quantifying the similarity, based on different comparison basis such as image intensities or age, and on different scales. However, as mentioned in the previous chapter, a high image correspondence should not be confused with high anatomical correspondence [65, 122]. Similarly, registration accuracy can usually only be quantified with surrogate measures [167]. Multiple surrogates were tested by Rohlfing and only label overlap scores of localised anatomical regions were able to provide a reasonable estimate. An accurate evaluation would require a dense set of landmarks, which is usually not available. In the next section we will provide an overview of different comparison strategies.

3.2 Comparison basis

Atlas selection is done on the basis of the comparison between the target image and the atlases. A number of criteria have been investigated in the literature. They can broadly be grouped into image-based and non-image based. Image-based criteria include features directly extracted from the given image or features indirectly extracted from image derivatives after processing the original. Examples for the former are image intensities or normalised image intensities, and for the latter, 3D surface meshes, registration consistency or deformation fields. Non-image-based comparison can be based on demographics such as age, sex or pathology.

3.2.1 Image intensities

The selection of the best candidates is most commonly based on the similarity of atlas and target intensity images after alignment [4, 5, 10, 212, 239, 240]. In general, methods based on nonlinear registrations of the atlas images to the target, e.g. [74], outperform methods based on affine registered images to the target, e.g. [64]. The former is computationally more expensive but provides better alignment, while the latter is more efficient but less accurate due to the poor alignment. Rikxoort et al. [217] presented a method with both characteristics. They efficiently aligned all atlases to the target with a fast affine registration, performed the ranking and selection, and used a computationally more expensive nonlinear registration only to align the selected atlases more accurately to the target.

3.2.2 Non-image information

The selection of the best candidates for a specific target can also be based on non-image information such as meta-data including age, sex or pathology. In [5], it was shown that age-based selection can achieve similar segmentation accuracy compared to the best results from intensity-based selection meth-

ods in young and middle aged groups. However, for the older aged group, age-based selection showed significantly better results. In a similar study, Aribisala et al. [10] investigated the impact of atlas selection for an older subject cohort. They found that the selection of different single atlases from the age-matched cohort barely had an impact on segmentation accuracy. However, the selection of a young adult brain atlas instead of age-matched atlases, led to systematic segmentation errors.

3.2.3 Registration consistency

Another atlas selection strategy was presented by Heckemann et al. [100], where the segmentation quality was measured based on two registration and transformation steps. A deformation field is estimated by registering an atlas- to the target image. This deformation field is applied to the corresponding atlas label to transform it to the target. The resulting transformed label is propagated back to the atlas image by applying to it the deformation field, estimated in a second registration between the target and atlas image. Registration and interpolation errors, introduced in this procedure, lead to differences between the original label and the forward-and-backward propagated label in the anatomical space of the atlas. These differences are measured with the Dice overlap coefficient and used as a quality measure. The atlases were ranked for each ROI according to their Dice overlap measures. A certain number of the highest ranked atlases were selected for fusion. Experiments were performed with the spline-based free-form deformation [177] with a control point spacing of 2.5 mm. Although they achieved higher Dice overlap scores compared to the use of randomly selected atlases, it should be noted that registration consistency does not provide a surrogate for measuring registration accuracy as shown by Rohlfing [167].

3.2.4 Anatomical geometry

An alternative atlas selection strategy for the segmentation of lymph node regions from CT scans was proposed by Teng et al. [204]. In contrast to other methods, the selection was based on landmarks extracted from 3D volumes of anatomical structures. The volumes were automatically segmented with a combined thresholding and active contouring method resulting in the delineated anatomical structure and a 3D surface mesh. Three different types of features were extracted from each anatomical landmark structure, including the structure's volume and its extent of the overall head and neck region, the vectors describing the structure's relative location to each other, and shape properties given as surface meshes. The feature vectors were used to create a weighted Euclidean distance matrix of the atlases to the target. One or multiple atlases with the shortest distances were selected and registered to the target. To achieve accurate registration, the 3D surface meshes were used as landmarks and incorporated as additional information in the registration between atlases and targets. The similarity measurement between the shapes of two meshes was calculated by the Hausdorff distance with smaller distances indicating a higher correspondence. The similarity ranking determined based on feature vectors was compared to the ranking determined by registering all atlases to the target. Results showed that 81% of the most similar candidates, selected based on features, matched the candidates, selected based on image registration, with a 96% chance of the top candidate to be within the top three subjects.

3.2.5 Deformation fields

The selection of the best candidates is most commonly based on image intensities, which has shown to provide a reliable basis for similarity measurement in high quality research scans. However, due to the non-quantitative image acquisition method used in MRI, image intensities are strongly influenced by the sequence, scanner, MRI system, coil and image reconstruction method.

Artefacts such as intensity non-uniformities, movement artefacts and partial volume effects can be introduced due to the scanner hardware or the subject. The presence of pathology further complicates image intensity classification and label affiliation. One approach, which has shown promising results for candidate selection, especially in the presence of pathology, is based on the comparison of deformation fields. The concept of using deformations for multivariate analysis has first been applied by Bookstein [37], who suggested the use of decomposed deformations, called principal warps, for multivariate statistical analysis. In the context of medical image analysis, the distribution of these principal warps allowed the discrimination and quantification of pathological pattern severity. The technique built the basis for the description of shape differences within or between subjects through their corresponding deformations. Thompson et al. [208] developed a framework to analyse the variability of cortical shape patterns and structural variations, and to classify deformity as normal variation or pathological abnormality. The deformation fields between subjects and an average population template in a normalised space were estimated to both bring them into correspondence, and carry information about their shape differences. Magnitude and direction of these local covariance tensors could be statistically analysed to provide probability values. A confidence region for the distribution of each anatomic point can be calculated in displacement space, indicating the likelihood of finding the cortical structure of interest at the specified anatomic location which, in turn, allows the characterisation of unusual anatomical positioning [206, 207]. The basic concept of classifying shape differences based on their deformation fields was employed by Commowick, Warfield and Malandain [60, 61] to select the best candidates for the segmentation of a target image. They hypothesise that the best result is obtained by selecting the most similar atlases to the target in displacement space, rather than intensity space, under the assumption that more similar images should have more similar deformations. Their objective function is to find the candidate $\widetilde{I}$ from the atlas images I_k, that requires the least amount of deformation $D_{T \to I_k}$ to nonlinearly register to the

target T as measured by the Euclidean distance $\|\|$:

$$\widetilde{I} = \arg\min_{I_k} \mathrm{d}\,(I_k, T) = \arg\min_{I_k} \|D_{T\to I_k} - \mathrm{Id}\| \tag{3.1}$$

While this approach requires the nonlinear registration of each atlas to the target, they proposed a computationally more efficient method. Instead of registering every atlas directly to the target, an average population template M was constructed offline. The resulting deformation fields $D_{M\to I_k}$ between each atlas I_k and M can be compared to the deformation field $D_{M\to T}$ estimated between the target and M at runtime. It is assumed that $T_{T\to I_k}$ can be composed by $T_{T\to I_k} \approx D_{M\to I_k} \circ D^{-1}_{M\to T}$:

$$d(I_k, T) = \left\|D_{M\to I_k} \circ D^{-1}_{M\to T} - Id\right\| = \sum_i \left\|(D_{M\to I_k} \circ D^{-1}_{M\to T})(i) - Id\right\| \tag{3.2}$$

3.3 Similarity metrics

Once a basis for comparison has been selected, a similarity or dissimilarity metric can be used to rank the atlases. While Wu et al. [168, 239] and Avants et al. [23] investigated the impact of the selection of the average population template and its construction on the segmentation accuracy, similar studies have been conducted for the selection of individual atlases for propagation to and segmentation of a target [1, 5]. The most commonly used metrics includind sum of squared differences (SSD), mutual information (MI), normalised mutual information (NMI) and correlation coefficient (CC) were compared. Aljabar *et al.* [5] nonlinearly registered all MR brain images to a common template space, where the similarity measurements with the above mentioned methods were conducted. Reference accuracy values were calculated by first ranking the atlas labels based on their Dice overlaps with the target labels and then fusing them into a final segmentation. It shall be noted that this is impossible in a real-life scenario where the target labels are

unknown. In their comparison NMI showed generally the best results. SSD was identified as the least reliable metric and CC and MI varied in between. In contrast, Acosta et al. [1] compared CC, SSD and MI on a set of non-linearly registered pelvic CT scans and identified SSD and CC as the most reliable. Although both studies used image intensities as a similarity basis, the use of different modalities, registration methods, fusion strategies and anatomical structures makes the comparison of these studies difficult. But it does outline that atlas selection and the choice of similarity measurement poses a challenge.

Another interesting study about the impact of the ranking on segmentation accuracy was presented by Ramus and Malandain [162]. They evaluated different ranking strategies independently from the number of selected candidates and atlas fusion method. The ranking strategies included CC, SSD, MI, and NMI applied to nonlinearly and affinely registered intensity-based methods, deformation-based methods, reference methods based on overlap and distance measures, and random ranking. All ranking methods were clustered and Spearmans rank correlation coefficients calculated to outline sub-groups of equivalent ranking methods, and the average correlation between the rankings of automatic methods and the ranking of the reference method. The resulting main clusters included the group of reference methods on one side of the spectrum and the random ranking on the other side of the spectrum. In between, intensity-based methods after nonlinear registration formed a group and intensity-based methods after affine registration formed a group with deformation-based methods. Results showed that the correlation between the reference and the clusters containing intensity-based methods after affine registration and deformation-based methods, was higher than between the reference and all clusters containing intensity-based methods after nonlinear registration. This suggests that deformation-based methods are more appropriate than any of the similarity metrics applied to intensity-based methods after nonlinear registration.

More complex similarity measurement methods, which have shown to be superior over conventional measurement methods, are based on distances between projections of atlas and target images in low-dimensional manifold space. In an evaluation by Duc et al. [75] three different manifold learning methods, including Isomap, Laplacian Eigenmaps and local linear embedding, were compared on a set of MR brain images with manually segmented hippocampus labels. The goals of the study were to find the best technique for atlas selection and the best choice of the manifold parameters. Local linear embedding with 11 dimensions reached the highest overlap measure, but they concluded that the choice of method and parameters should be determined empirically, since it also depends on the dataset and the number of atlases used. Overall results reached state-of-the-art accuracy or exceeded results obtained with conventional selection methods.

3.3.1 Basic similarity metrics

The SSD is ideally used if the difference between the two images is only Gaussian noise, because of its sensitivity to outliers such as intensity inhomogeneities. Consequently, without further intensity mapping, it should only be used for images of the same modality with an identity relationship between their intensity ranges. For two images, a target T and an atlas A_M, with N voxels the SSD can be calculated as

$$SSD = -\frac{1}{N}\sum_{x\epsilon\Omega} |T(x) - A_M(x)|^2 \tag{3.3}$$

Similarly, CC can only be used for mono-modal image comparison. In contrast to SSD, the intensity ranges of the images can differ but require a linear relationship. The CC of two images with their respective average intensities μ_T and μ_{A_M} can be calculated as

$$CC = \frac{\sum_{x \epsilon \Omega} (T(x) - \mu_T)(A_M(x) - \mu_{A_M})}{\sqrt{\left(\sum_{x \epsilon \Omega} (T(x) - \mu_T)^2\right)\left(\sum_{x \epsilon \Omega} (A_M(x) - \mu_{A_M})^2\right)}} \tag{3.4}$$

for every voxel x.

A method less sensitive to different intensity ranges, which makes it suitable for the comparison of images of different modalities, is MI or NMI. Its goal is to find the amount of shared information in two images. The amount of shared information in perfectly aligned images is a minimum and increases the more dissimilar they are. For instance, the difference image of two perfectly aligned images is a uniform image with zero entropy, which increases with misregistration. These information theoretic techniques compute the similarity of two images from the frequency of their corresponding joint histogram. MI of an atlas image A and a target T can be expressed as

$$MI = H(A) + H(T) - H(A, T) \tag{3.5}$$

with $H(X)$ being the marginal entropy of an image X, given by

$$H(X) = -\sum_x p_X(x) log(p_X(x)), \tag{3.6}$$

the joint entropy, which is calculated from the probabilities of pairs of image values occurring together, given by

$$H(A, T) = -\sum_a \sum_t p_{A,T}(a, t) log(p_{A,T}(a, t)) \tag{3.7}$$

and with $p_{A,T}$ describing the joint probability distribution of a voxel associated with images A and T. To make MI more robust, normalised mutual information (NMI) has been introduced and can be calculated with

$$NMI = \frac{H(A) + H(T)}{H(A, T)} = \frac{MI(A, T)}{H(A, T)} + 1 \tag{3.8}$$

3.3.2 Manifold learning

With a growing amount of population-based MRI studies, machine learning and pattern recognition algorithms have become important tools in computational neuroscience. These data driven approaches encompass filtering algorithms, measures for determining coherence or relations in the data and classification strategies. One class of methods of particular interest for medical image analysis and especially for registration, segmentation and classification is called manifold learning and facilitates dimensionality reduction and projection techniques [6].

Conventional similarity measurement methods from the previous section operate in high-dimensional feature space with as many dimensions as there are features, where the number of features is determined by the comparison basis. In contrast, manifold learning techniques allow the projection of data from a high to lower dimensional representation while respecting the intrinsic geometry of the data. Applied to medical images and image intensity as the similarity basis, this can be understood as follows. An image with 8 million voxels can be represented as one point in a space with 8 million dimensions, with coordinates given by the intensity values of each voxel. Due to similarities in images of a certain population, their representations in this high dimensional space can be seen as a cloud of points lying very close to each other. This means that many fewer dimensions might be necessary to represent this certain group of images, allowing these points to be embedded in a low dimensional or sub-manifold in the high dimensional space. Manifold embedding methods are able to find patterns in the data based on similarities and differences in the data and can thereby significantly reduce the amount of necessary dimensions for their representation and further processing. In the context of atlas selection this has some important implications. As mentioned previously, similarity is conventionally quantified with some type of metric, which is capable of measuring the distance between two points in the high-dimensional space. For instance, SSD is the sum of squared differ-

ences between the respective coordinates of two images in each dimension. Considering the images are close to or lie on a lower-dimensional manifold, Euclidean-based distance measures in high-dimensional space are not always a meaningful representation of similarity. In contrast, the use of the geodesic distance, calculated on the manifold structure, is more representative and can be approximated by the Euclidean distance in a lower dimensional manifold space. An important property of manifold learning methods is their ability to preserve the intrinsic geometry of data when projecting them from high-dimensional space into lower-dimensional manifold space. In the last two decades, different manifold learning strategies have been developed, each with the goal to preserve a different geometrical property by optimising some objective function [6, 45, 75].

In summary, nonlinear methods such as Isomap and Locally Linear Embedding (LLE), compared to linear methods such as Principal Component Analysis (PCA), are able to find more complex patterns in the data and preserve the intrinsic geometry [89, 96]. Although these are huge advantages over linear methods, they do not come without drawbacks. The need for the initial generation of a graph, connecting all data points, is computationally expensive, considering the high dimensional space of about 24 million dimensions in case of deformation fields. In contrast, PCA is only based on second order statistics, given by the covariance matrix, which makes them more feasible in some real world applications [110, 202, 234]. Overall, it has been shown that manifold embedding methods can reduce the dimensionality, while still preserving the geometry of the data, and allow more accurate similarity measurement between images on a low-dimensional manifold embedded.

3.3.2.1 Principal component analysis (PCA)

PCA [112] is a classic method to highlight similarities and differences by identifying patterns in high dimensional data. An efficient implementation for the application to images was proposed by Turk and Pentland [213], where

the goal is formulated as finding the principal components of the images that account for the largest variation. Mathematically it can be explained as finding the eigenvectors and the corresponding eigenvalues of the covariance matrix of the images. By sorting the components according to their eigenvalues in a decreasing order, a set of dimensions can be identified whereby the first principal component represents the dimension that accounts for the largest variance in the data, the second for the second largest variance and so on. The number of dimensions can be reduced by eliminating redundant information and ignoring the components that account for the least amount of variance, without losing too much information. The original images and new targets can be projected into this eigenimage space, where they can be compared and reconstructed from the eigenimages. Because redundancy is measured by correlations, a drawback of this method is its dependency on only second order statistics. Consequently, it lacks information for higher order statistics and is therefore classified as a linear dimensionality reduction method.

Given a set of M mean-corrected images in matrix form $A = [\mathbf{\Phi}_1, \mathbf{\Phi}_2 ... \mathbf{\Phi}_M]$, the covariance matrix can be calculated as

$$C = \frac{1}{M} \sum_{n=1}^{M} \mathbf{\Phi}_n \mathbf{\Phi}_n^T = AA^T. \tag{3.9}$$

With large images, finding the eigenvectors of C can become computationally expensive. The efficient method presented by Turk and Pentland [213] assumes that the number of images is smaller than the number of dimensions, which allows the calculation of the eigenvectors $\mathbf{v}_i$ of the much smaller matrix $A^T A$ such that

$$A^T A \mathbf{v}_i = \lambda_i \mathbf{v}_i. \tag{3.10}$$

To determine the eigenvectors of C, both sides can be multiplied by A which leads to

$$AA^T A\mathbf{v}_i = \lambda_i A\mathbf{v}_i \tag{3.11}$$

where $A\mathbf{v}_i$ are the eigenimages $\mathbf{u}_l$ of $C = AA^T$. A mean-corrected image $\mathbf{\Phi}$ can be projected into eigenimage-space by

$$w_k = \mathbf{u}_k^T \mathbf{\Phi} \quad \textit{for } k = 1,\ldots,M' \tag{3.12}$$

where w_k represents the contribution of the k-th eigenimage in the reconstruction of the target. The number of eigenimages M' for the projection is variable, with more eigenimages accounting for more of the variance.

3.3.2.2 Isomap

A method capable of discovering nonlinear degrees of freedom in the data, while preserving their intrinsic geometry, is the manifold embedding technique Isomap [203].

The goal of the algorithm is to find a low-dimensional representation of the data, that best preserves the distances between the data-elements measured in high-dimensional space. This can be exemplified by comparing geodesic and Euclidean distances between two data points. For example, if we consider a spiral of data points in 3D, two points might be close to each other in terms of their Euclidean distance but far apart, following the embedded manifold (Fig. 3.1). In order to capture this nonlinearity, a distance matrix, which contains the geodesic distances between each of the data items, is created. This is done via a neighbourhood graph that connects each data item to a specified number of closest neighbours, as measured by Euclidean distances in high-dimensional space. Then the shortest paths between all pairs of elements of the graph can be estimated and a lower-dimensional embedding generated. Similarly to PCA, the embedding is constructed by calculating the principal components and projecting the data points onto it. The objective function of Isomap is:

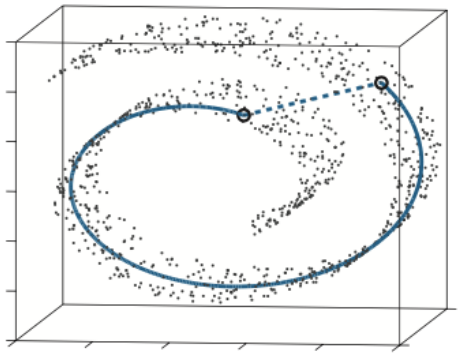


Figure 3.1: The distance on the manifold (solid blue line) differs from the Euclidean distance (dotted blue line) [203].

$$\Phi(Y) = \sum_{i=1}^{n}\sum_{j=1}^{n}(g_{ij}^2 - \|y_i - y_j\|^2) \tag{3.13}$$

with g_{ij} representing the geodesic distance between two data points, e.g. atlases a_i and a_j from a set of n atlases $A = (a_i, ..., a_n)\epsilon\mathbb{R}^D$ in a high-dimensional space. $Y = (y_1, ..., y_n)\epsilon\mathbb{R}^d$ is the new transformed dataset in the low-dimensional space.

3.3.2.3 Local linear embedding (LLE)

Another unsupervised learning algorithm for the computation of low-dimensional, neighbourhood-preserving embeddings of high dimensional input data is LLE [175]. While Isomap aims to preserve the geodesic distances between pairs of data points in high-dimensional input space, which requires the pairwise geodesic distance measure between all points, LLE represents the local geometry of patches by reconstructing each data point from linear coefficients of its neighbours. It is assumed that there is a sufficient number of data points, so that each point is close to its neighbours and a patch on the manifold. Due to this local reconstruction, pairwise distance measurements outside the respective local patch size are not required. The objective function

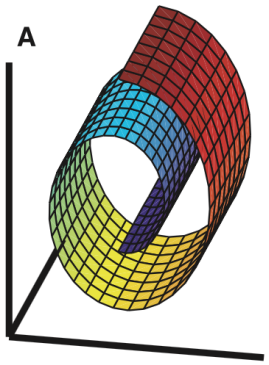


Figure 3.2: Data points and patches on manifold in 3 dimensions [175].

$$\Phi(Y) = \sum_{i=1}^{n} \left\| y_i - \sum_{j \epsilon N_k(i)} w_{ij} y_{ij} \right\|^2 \tag{3.14}$$

measures the cost presented by the reconstruction error and is minimised by solving a least-squares problem with constraints on the weights of each of the neighbours' contributions. The weights w_{ij} represent the contribution of data point j to reconstruction i where data point j has to be within the neighbourhood and the weights sum up to one. This leads to symmetry for any particular data point and its neighbours, making them invariant to rotations, rescaling, and translations, which preserves the intrinsic geometry.

3.4 Our approach to TST building

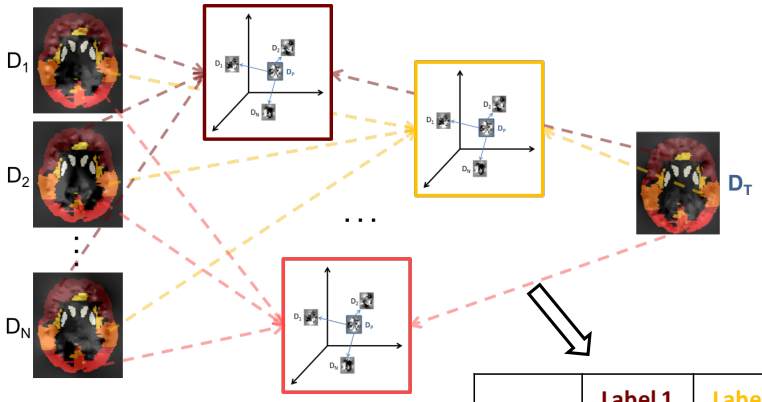
We build our TST image from the atlas images combined in such a way as to satisfy the following two constraints: (a) the TST should be as similar to the target as possible, to make it easier to estimate an accurate non-linear mapping between them and (b) the number of composed deformation fields should be limited to reduce the risk of an imprecise transformation.

In our approach we employ an average population template for computational efficiency (Chapter 2) and construct a TST locally similar to the target image by working at the label level, merging together regions from those atlas

images selected for their similarity to the corresponding region in the target. Compared to other methods, our similarity measurement is based on the L^2 distance in nonlinear manifold embeddings constructed from the regions of deformation fields and the TST is constructed from the warped atlases in the space of the target. This has the following advantages: (1) using deformation fields rather than image intensities makes for a similarity measure much less influenced by the intensity artefacts and inhomogeneities frequently found in MR images, (2) using label-specific nonlinear manifolds makes it possible to better take into account the local geometry of the atlas and target images and (3) assembling the TST from parts of the warped atlas intensity images in the space of the target makes for an even more similar TST. The TST serves as an intermediate template and registering the target to the TST is expected to refine a potentially large, difficult registration between the target and average population template. This adds one nonlinear registration at runtime and, in turn, one deformation to the composition. However, this registration is performed directly, rather than via intermediate images, and between two images very similar to each other, the TST and the target. The atlas labels can be propagated to the target by composing the deformation fields between the corresponding atlas and the average population template, between the population template and the TST and between the TST and the target. Consequently, the number of composed deformation fields is constant and does not depend on the number of atlases.

3.4.1 Manifold embedding

For each label r we build a non-linear manifold space from the corresponding region of the deformation fields estimated between the atlas images and the population template, and between the target and the population template. This region, B_r, consists of the minimum set of all voxels with non-zero probability across the individual label maps $\{\widetilde{L_j^r}\}_r$, i.e. $B_r = \{x | \exists j \in [1, N], \widetilde{L_j^r}(x) > 0\}$. Note that this makes for potentially overlapping regions



	Label 1	Label 2	...	Label R
Atlas 1	λ_{11}	λ_{12}		λ_{1R}
Atlas 2	λ_{21}			
...				
Atlas N	λ_{N1}			λ_{NR}

Figure 3.3: A manifold is built for each label from the corresponding ROI of the deformation fields. The distances in manifold space provide a measure of dissimilarity.

across labels, which has to be taken into account when assembling the TST and fusing the labels. Following the recommendation of Duc et al. [75], we considered and empirically tested both LLE and Isomap and selected the latter as it exhibited the best overall results. The parameters were fine-tuned by systematically varying the number of neighbours and dimensions, as they did.

We extract the set of displacement vectors in the region from both the atlas deformation fields, $\{D_{\widetilde{I}_j \to \bar{I}}\}_{j=1...N}$, and from the target deformation field, $D_{\widetilde{T} \to \bar{I}}$. The resulting sets of displacement vectors, $\{U^r_{\widetilde{I}_j} = \{D_{\widetilde{I}_j \to \bar{I}}(x) | x \in B_r\}\}_{j=1...N}$ and $U^r_{\widetilde{T}} = \{D_{\widetilde{T} \to \bar{I}}(x) | x \in B_r\}$ respectively, are then mean corrected and we compute their pairwise L^2 distances to serve as input for the Isomap algorithm. For each label, we get a projection of $U^r_{\widetilde{T}}$ and of the $\{U^r_{\widetilde{I}_j}\}_j$ onto the lower dimensional manifold space (Fig. 3.3). We can then calculate the L^2 distances between the target label projection and the atlas label projections and rank them. Overall, we get a distance matrix, Λ, with

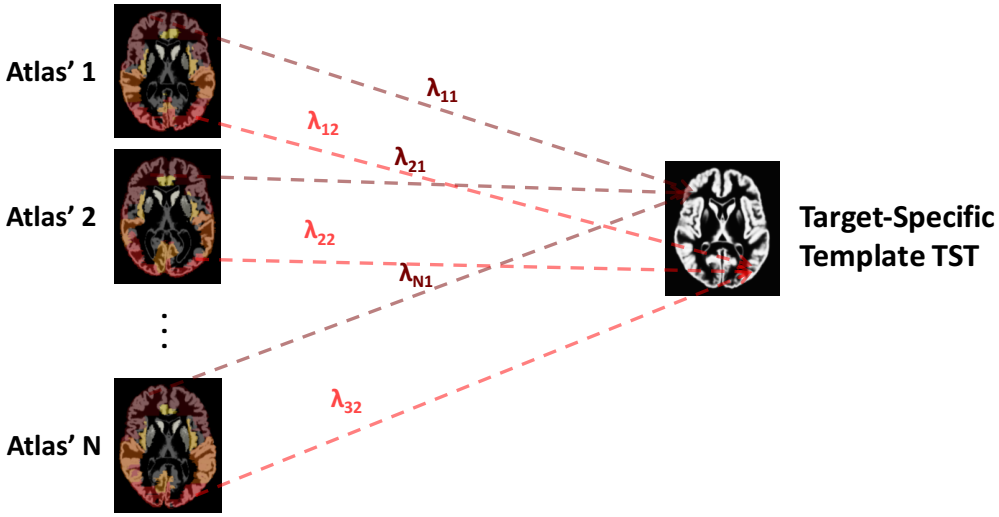


Figure 3.4: The distances in manifold space are used as weights for the construction of the TST from regions of the warped atlas images.

R columns, one per label, and N rows.

3.4.2 TST construction

For each label r we warp the atlas intensity images $\widetilde{I_{j_i}}$ on the target $\widetilde{T}$ by composing $D_{\widetilde{I_{j_i}}\to\bar{I}}$ and $D_{\widetilde{T}\to\bar{I}}{}^{-1}$. This satisfies constraint (b) and results in a set of warped intensity images $\widehat{I_{j_i}}$, from which we extract the intensities in region B_r. Here we use the normalised manifold distances as weights for the candidate segmentations: $\omega_i^r = 1 - \frac{\lambda_i^r}{\sum_j \lambda_i^r}$ where $\sum_{j=1}^N (w_j^r) = 1$ to compute a weighted sum $H^r = \{\sum_{j=1}^N \omega_j^r \cdot \widehat{I_{j_i}}(x) \; | x \in B_r\}$ (Fig. 3.4). Finally, we assemble the TST by putting together the $\{H^r\}_r$. Two special cases can occur: when regions overlap, we average the intensity values of the corresponding voxels H^r and for regions that are not part of a label we use the intensity value of the target image.

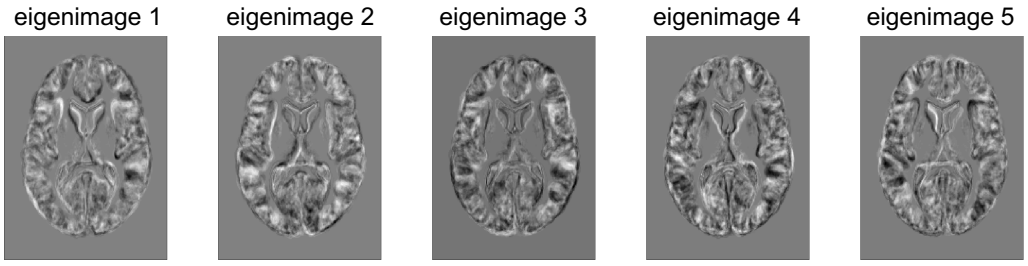


Figure 3.5: The five highest ranked eigenimages in decreasing order.

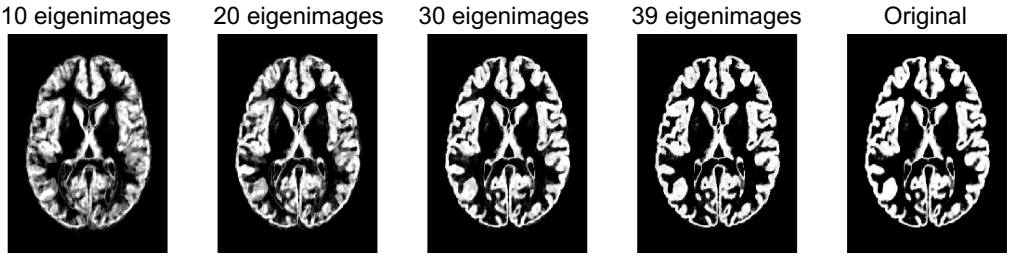


Figure 3.6: Reconstruction of an atlas image from the training set with an increasing number of eigenimages.

3.5 Experiments

3.5.1 Reconstruction of a GM image slice with eigenimages

The 40 images of the LONI dataset were skull-stripped, affinely registered and classified into their tissue maps. The same axial centre slice was extracted from all GM tissue maps resulting in 40 2D images. 39 images were used as atlases for training and one image as target for testing. As outlined in Section 3.3.2.1, the eigenimages were determined from the atlases and a variable number of them used to reconstruct an atlas from the training set, to reconstruct the left-out target and to find the most similar atlas to the target image.

The eigenimages, illustrated in Fig. 3.5 were calculated from the training set and ranked in decreasing order of their corresponding eigenvalues. A random atlas image from the training set was reconstructed with the 10, 20, 30

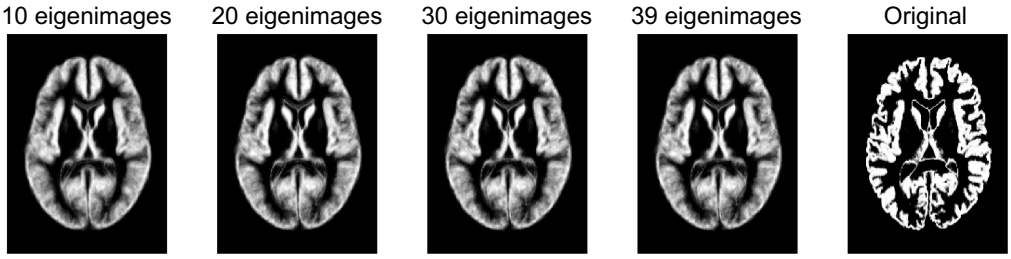


Figure 3.7: Reconstruction of a non-atlas target image with an increasing number of eigenimages.

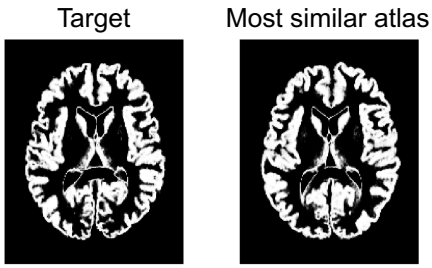


Figure 3.8: Target and the most similar atlas image identified in the space of eigenimages.

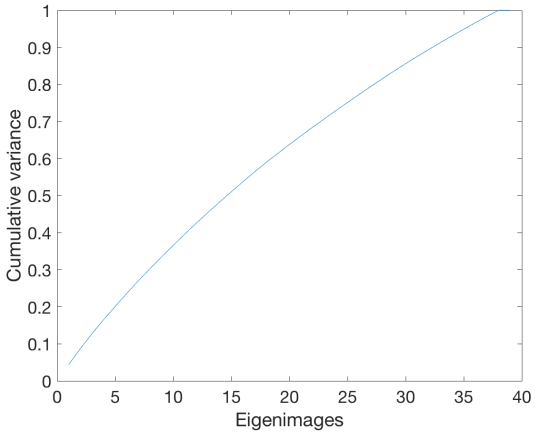


Figure 3.9: Cumulative variance explained by the eigenimages.

and 39 highest ranked eigenimages, which account for ∼38% of the variance (Fig. 3.9). The reconstruction with 10 eigenimages showed poor quality and similarity to the original but with an increasing number of eigenimages a

higher degree of detail and similarity to the original image was added (Fig. 3.6). A perfect reconstruction was achieved with all 39 eigenimages, which accounts for 100% of the variation in the training set. An acceptable reconstruction can usually be achieved with eigenimages accounting for a variance of at least 65%. The reconstruction of a target image not in the atlas set showed only little improvement with the use of more eigenimages (Fig. 3.7). One reason might be the large morphological inter-subject variability shown in MR brain images and the coarse alignment of corresponding anatomical structures. Consequently, the high variation can not be captured by the small number of 39 atlases and, in turn, eigenimages.

3.5.2 Reconstruction of a deformation field with eigenimages

The 40 images of the LONI dataset were pre-processed as outlined in Section 2.6. The population template was created from the tissue maps of 30 randomly selected images. Each of the 10 remaining images was nonlinearly registered to this same template. The corresponding deformation fields, estimated during the registration process, were used as atlases for training and targets for testing respectively. The eigenimages were determined from the atlas set and a variable number of them used to reconstruct an atlas from the training set, reconstruct a left-out target and find the most similar atlas to a target. The inverse of the reconstructed deformation field was applied to the population template to obtain the individual target tissue map and compared to the original.

The eigenimages were calculated from the training set and ranked in a decreasing order according to their corresponding eigenvalues (Fig. 3.10). Similarly to the results from the previous experiment, an atlas from the training set was reconstructed with the 20 highest ranked eigenimages, which account for over 70% of the variance in the training set (3.14). The reconstruction showed a high level of similarity to the original (Fig. 3.11), which could be

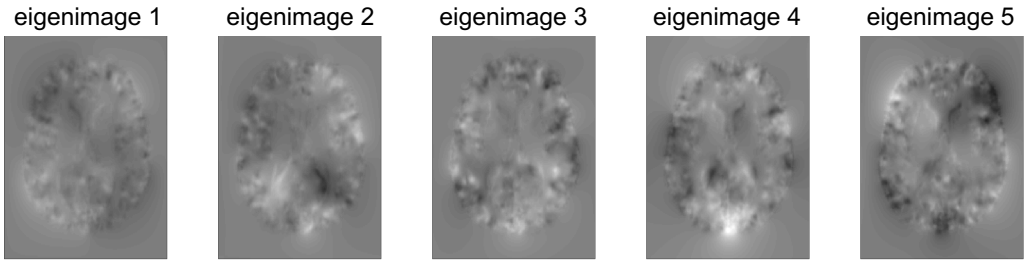


Figure 3.10: The five highest ranked eigenimages in decreasing order.

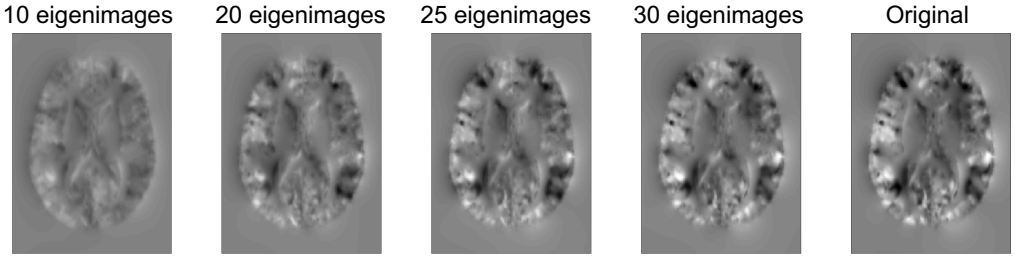


Figure 3.11: Reconstruction of an atlas from the training set with an increasing number of eigenimages.

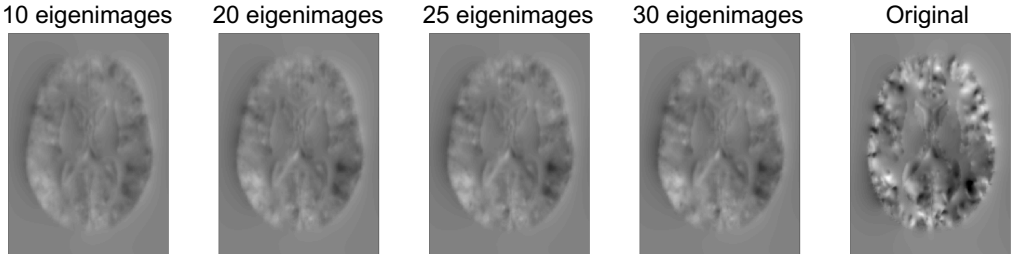


Figure 3.12: Reconstruction of a target with an increasing number of eigenimages.

further improved until a perfect reconstruction was achieved with all eigenimages. The reconstruction of a left-out target shared only local similarities with the original (Fig. 3.12) and the use of more eigenimages showed only incremental improvement. This is also reflected by the atlas identified as the most similar to the target, which shows only a limited degree of similarity (Fig. 3.13). Local differences become even clearer when applying the inverse of the reconstructed deformation field to the population template,

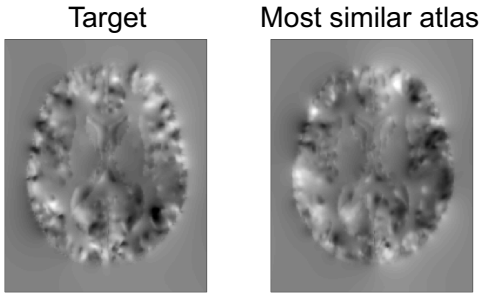


Figure 3.13: Target and the most similar atlas identified in the space of eigenimages.

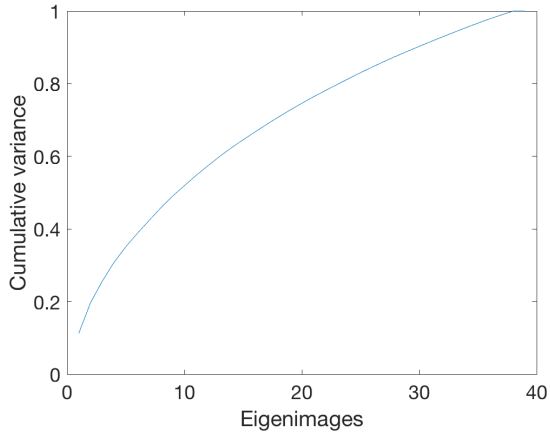


Figure 3.14: Cumulative variance explained by the eigenimages.

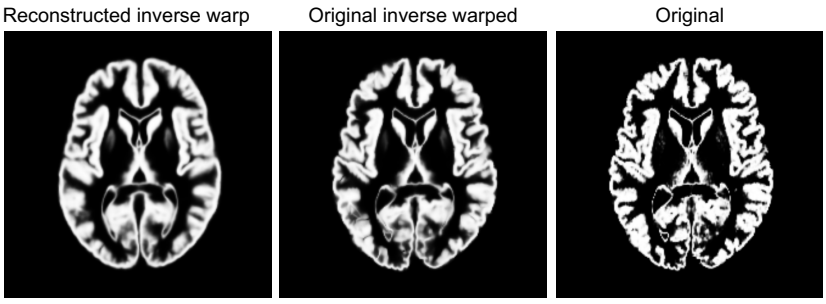


Figure 3.15: The target obtained by applying the inverse of the reconstructed deformation field to the population template (left), the target obtained by applying the inverse of the original deformation field to the population template (middle), and the original image as reference (right).

yielding the individual target tissue map (Fig. 3.15 left). This reconstructed individual target tissue map is blurrier and anatomical structures differ considerably from the target image (Fig. 3.15 right) and the individual target image retrieved by applying the inverse of the original deformation field to the population template (Fig. 3.15 middle).

Results suggest that a global approach, where an image or deformation field is reconstructed as a whole, is not sensitive enough to capture the local anatomical variability. Since PCA is not capable of detecting complex patterns in the data, more sophisticated nonlinear manifold embedding methods should be considered.

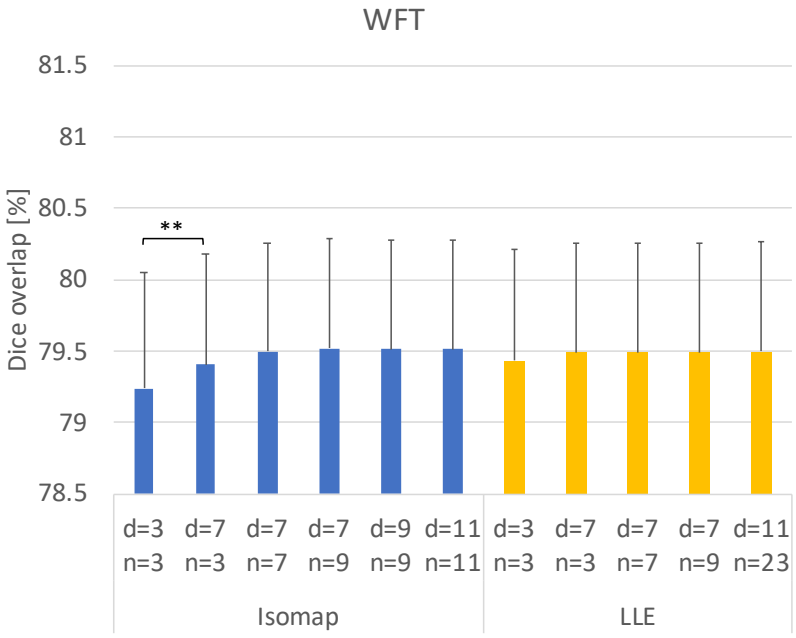
3.5.3 Comparison of nonlinear dimensionality reduction methods and parameter optimisation

The images of the LONI dataset were processed as outlined in Section 2.6. We used 53 labels of 30 atlases and 5 target images with low, medium and high expected segmentation accuracy. The two manifold learning methods LLE and Isomap were used for determining the ranking of the most similar atlases and weights for label fusion. Both methods were parametrised with varying numbers of dimensions and neighbours. The label fusion was performed with two different fusion strategies, weighted fixed tresholding (WFT-1/sim) and weighted intensity thresholding (WIT-1/sim), which are explained and evaluated in Sections 4.9-4.10 in more detail.

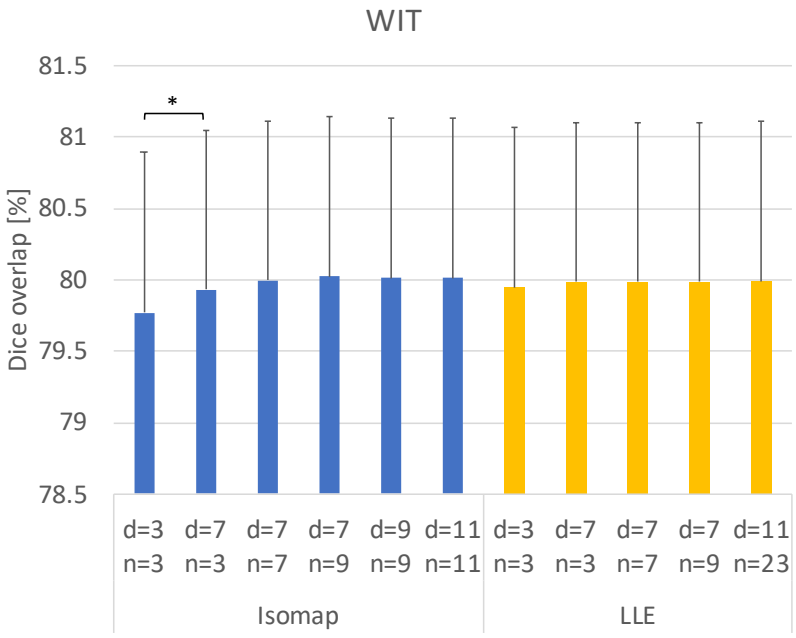
Since the main focus of this section is on the impact of using a TST we will only provide a brief introduction of the fusion methods. Both methods assign a local weight (W) to each label of an atlas candidate. Each weight is derived from the distance between the projection of its corresponding atlas ROI and target image in manifold space. We use two different ways of calculating the weights, which will be referred to as 1/sim and exp. The weighted sum of the probabilistic labels provide an estimate of the final segmentation. The final result was obtained by binarising this estimate either purely based

on the probabilistic values and a fixed threshold (FT) or by also considering the intensity values of the target image in the ROI (IT). For the latter, we check whether the intensity value of a voxel falls within the intensity distribution estimated from high probability regions of the fused candidate labels in target space. Both binarisation methods can also be applied to unweighted (U) probabilistic estimates. This leads to multiple combinations, that are unweighted fixed tresholding (UFT) and intensity thresholding (UIT), and weighted fixed thresholding (WFT-1/sim or WFT-exp) and intensity thresholding (WIT-1/sim or WIT-exp). Majority voting (MV) counts the number of votes, each voxel receives from the candidates and assigns the label with the most votes to the voxel. A voxel gets a vote from a candidate when the probability of finding the label at the location is larger than 0.5. In the following experiments the weighted fusion method was used when a TST was constructed and unweighted fusion or majority voting was used when no TST was constructed, since the weights are usually only available when we construct a TST. Local thresholding was used for validation on the IBSR dataset, where the binarisation is based on a locally determined threshold. We also provide the results when patch-refinement was used for the NIREP dataset to allow a direct comparison.

For both fusion strategies and manifold learning methods an improvement in segmentation accuracy was observed with an increasing number of dimensions, which flattened out after 7 dimensions and 7 neighbours (Fig. 3.16). The improvement was mainly caused by the increasing number of dimensions, while the number of neighbours had little influence on the accuracy with Isomap and no influence with LLE. These findings are in line with Duc *et al.'s* more comprehensive analysis [75], in which they found that improvement is mainly governed by the change in number of dimensions rather than neighbours. In their study on hippocampus labels the maximum segmentation accuracy was reached with LLE and 11 dimensions and 23 neighbours on a dataset of 110 atlases, but it was pointed out that the choice of manifold embedding strategy and its parameters might differ based on the anatomical



(a) Weighted Fixed Thresholding



(b) Weighted Intensity Thresholding

Figure 3.16: LLE and Isomap were tested with varying numbers of dimensions d and neighbours n for the construction of the low-dimensional manifold embedding. Error bars represent the standard deviation. A paired t-test was used to assess statistical significance: one asterisk represents $p < .05$, two asterisks represent $p < .01$, and three asterisks represent $p < .001$

structure and size of the dataset. In our experiments these settings showed no improvement, which might be caused by the smaller number of atlases and the varying size of the anatomical ROIs in the dataset. Interestingly, most studies have used 2 or 3 dimensions to allow simple visualisation of the connected manifold graphs, while the use of 3 dimensions has shown to perform worst in our study. For all following experiments we chose 7 dimensions and 7 neighbours, which has shown to provide robust results with both manifold embedding strategies and fusion methods. It also represents a reasonable choice, considering the sizes of our datasets where the smallest consists of only 13 atlases.

3.5.4 Method validation on the LONI dataset

In a cross-validation experiment the LONI dataset was divided into 4 subsets, each consisting of 30 atlases and 10 left-out target images. Isomap was used with the parameter settings from the previous section for the manifold learning. A visual comparison of the TST, population template and target, and their respective deformation fields between target and TST and between target and population template was performed. The overall accuracy gain by using a population template and our TST was compared to using only the population template for label propagation. The fusion for the latter method without using a TST was performed with majority voting (MV), where each voxel is assigned the label with the most occurrences in the candidate labels (Section 4.1), and unweighted fixed thresholding (UFT). When using the TST, the fusion was performed by weighted fixed thresholding, where the weights were calculated with two methods WFT-1/sim and WIT-1/sim (Section 4.9). For a more comprehensive evaluation of different fusion methods please see Section 4.10. A paired t-test was used to assess statistical significance.

Figure 3.17 shows the same axial 2-D view cut through the 3-D grey matter tissue map of a participant from the LONI dataset, the LONI grey matter

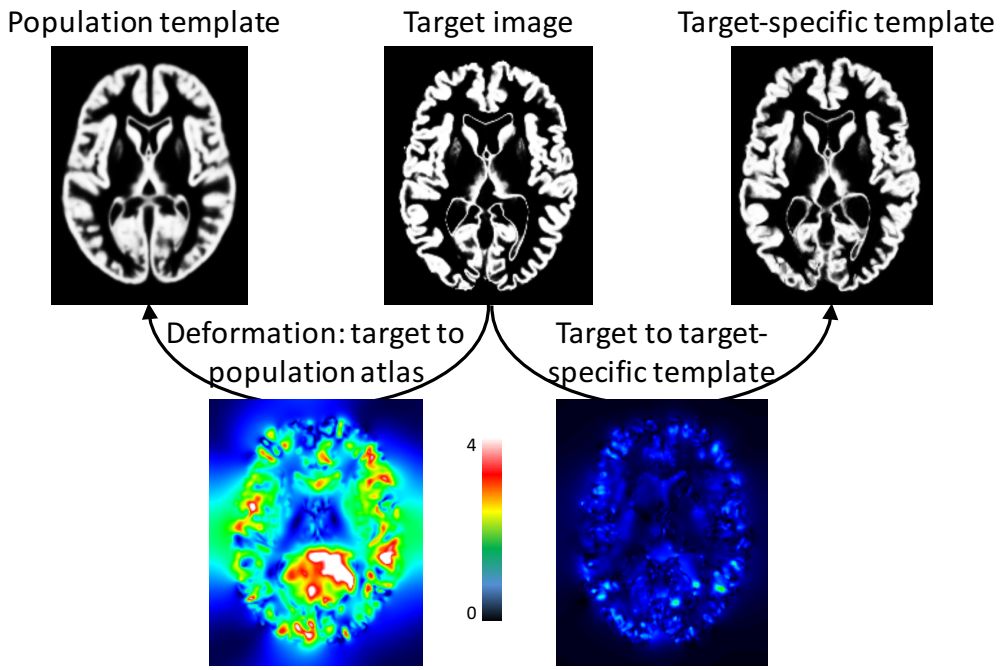


Figure 3.17: The same axial 2-D view cut through the 3-D gray matter tissue map of a subject. The TST is much more similar to the target than the population template as evidenced by their corresponding deformation fields [13].

template, the participant's gray matter TST and the calculated deformation fields. Note how the TST is more similar to the target image than the population template $\overline{I}$, as evidenced by the visual comparison of the GM tissue maps. The magnitude of the deformation fields, estimated between the target and the population template, $D_{\widetilde{T}\to\overline{I}}$ (bottom left) and between the target and the TST, $D_{\widetilde{T}\to\mathrm{TST}}$ (bottom right) shows that only a small deformation is required to align the the TST and target. The use of a TST can reduce the likelihood of getting stuck in local minima during the registration and consequently, the occurrence of registration errors.

Our best Dice overlap of 80.53 $\pm$ 1.17% over all targets and labels was achieved with a TST and our weighted voting and intensity thresholding method (Fig. 3.18). Paired samples t-tests showed significant increase in accuracy between MV (79.14 $\pm$ 1.24%) and UFT (79.81 $\pm$ 1.07%) without a

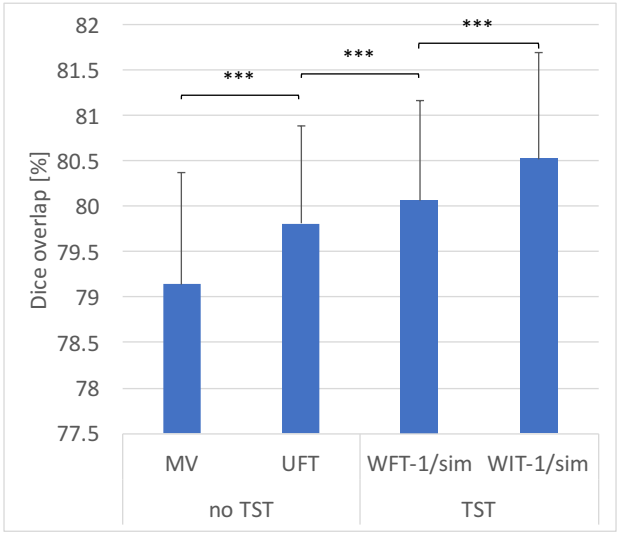


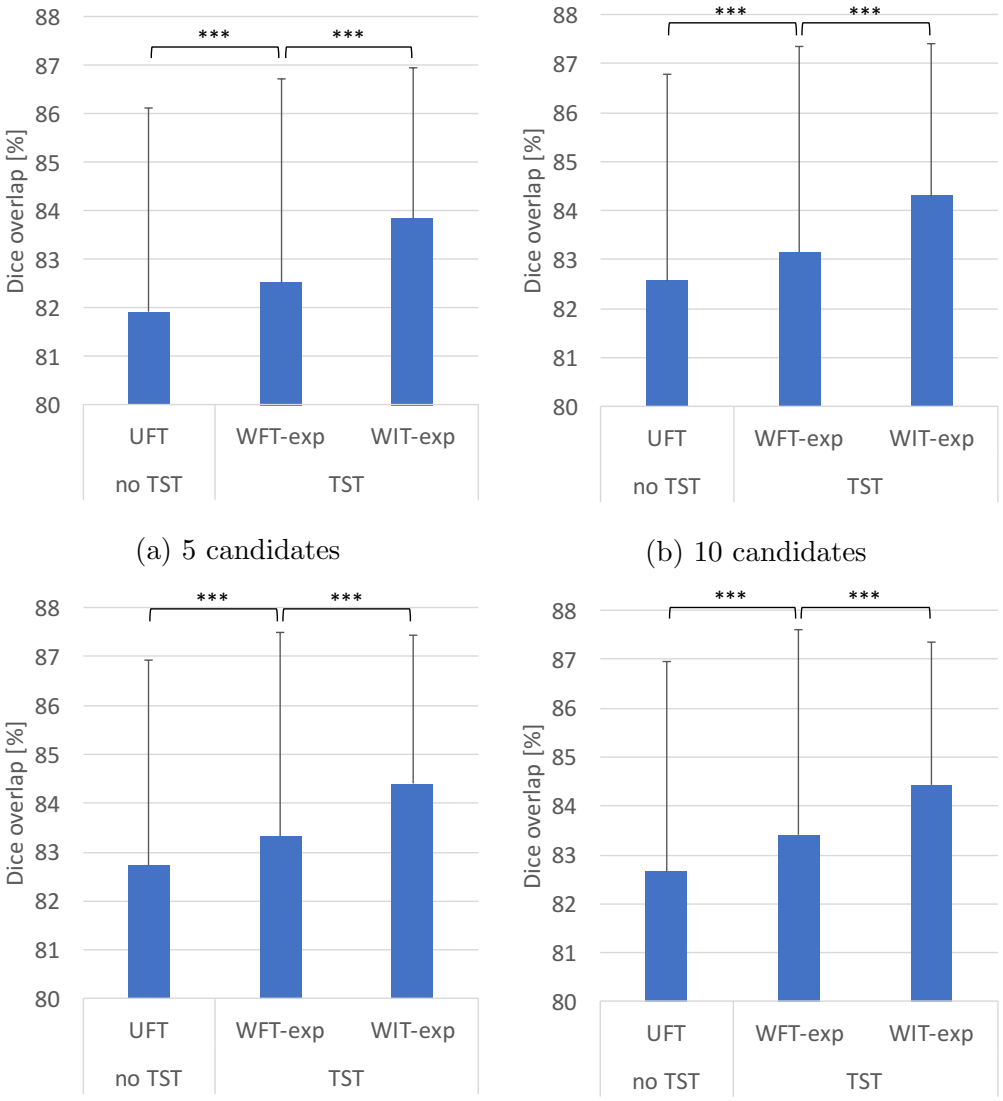
Figure 3.18: Improvement in segmentation accuracy by using our TST on the LONI dataset. Error bars represent the standard deviation.

TST ($p = 1.6 \times 10^{-8}$) between UFT without a TST and WFT-1/sim (80.07 $\pm$ 1.1%) with a TST ($p = 3.1 \times 10^{-20}$) and between WFT-1/sim and WIT-1/sim (80.53 $\pm$ 1.17%) with a TST ($p = 5.3 \times 10^{-10}$).

3.5.5 Method validation on the ADNI-HarP dataset

The ADNI-HarP dataset was divided into 5 subsets of 26 target images and 105 atlases each. Four images were excluded from the set due to problems with the orientation information. In a cross-validation experiment the Dice overlap was measured after direct label propagation and fusion without a TST and compared to the overlap when using a TST. Fusion for the former was performed with simple UFT and for the latter with WFT and WIT. The fusion was performed with 40, 15, 10 and 5 candidates. Note that although no TST was used for the label propagation for the former, the manifold embedding and the ranking and selection of the most similar candidates was performed.

Segmentation accuracy when using a TST and the associated weights for fusion improved for all four candidate choices, compared to propagation



(a) 5 candidates

(b) 10 candidates

(c) 15 candidates

(d) 40 candidates

Figure 3.19: Improvement in segmentation accuracy by using our TST, different fusion methods and candidate choices on the ADNI-HarP dataset. Error bars represent the standard deviation.

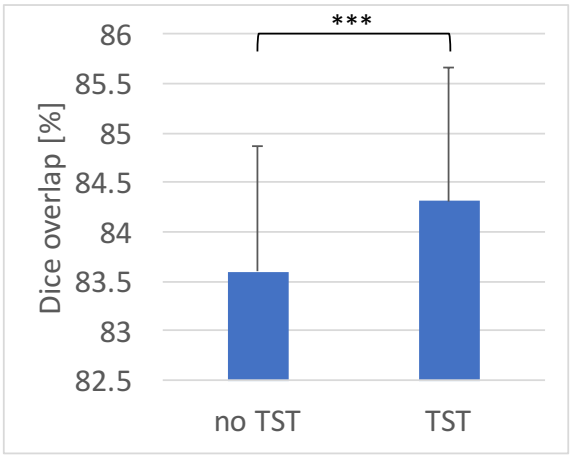


Figure 3.20: Improvement in segmentation accuracy by using our TST and local thresholding method on the IBSR dataset. Error bars represent the standard deviation.

without TST (p<.001) (Fig. 3.19). The intensity-based thresholding method WIT-exp showed superior results to fixed thresholding WFT-exp when a TST was used (p<.001).

3.5.6 Method validation on the IBSR dataset

In a cross-validation experiment the IBSR dataset was divided into 6 subsets of 3 images. Each fold was used as a target image set with the remaining images as atlases. The CSF label was removed as it was shown that its use leads to a strong bias in performance and accuracy [215], leaving a total of 31 labels. We compared segmentation accuracy when propagating atlas labels with our TST to label propagation without the TST. We used weighted fusion for the former and unweighted fusion for the latter and local thresholding for both methods (Section 4.9). Segmentation accuracy could be significantly improved from 83.6 ± 1.26% without a TST to 84.31 ± 1.35% with a TST ($p = 4.9 \times 10^{-5}$)(Fig. 3.20),

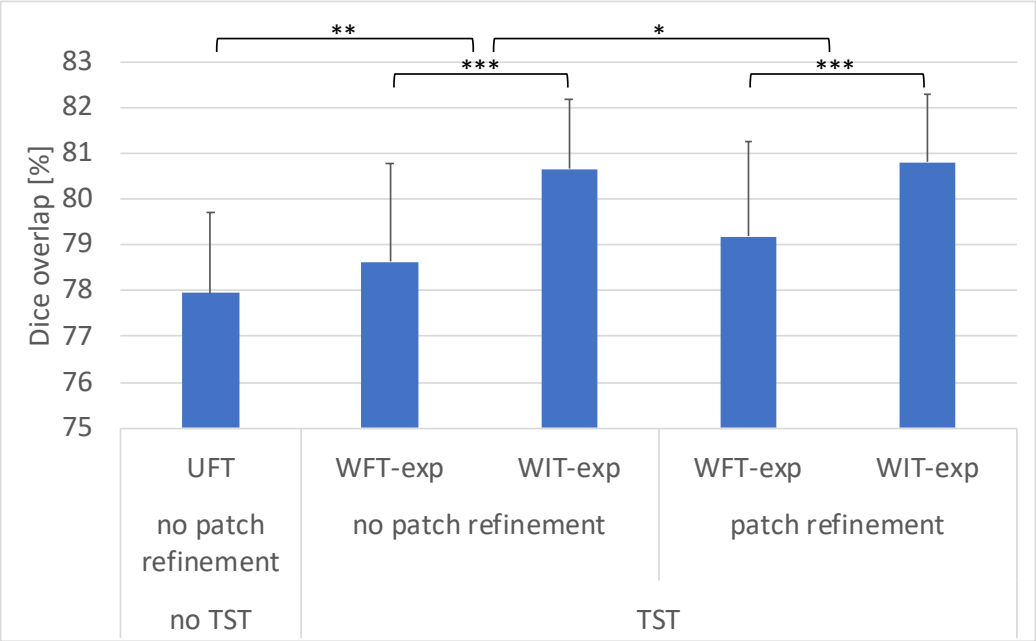


Figure 3.21: Improvement in segmentation accuracy by using our TST, different fusion methods (WFT-exp, WIT-exp) and patch-based refinement on the NIREP dataset. Error bars represent the standard deviation.

3.5.7 Method validation on the NIREP-NA0 dataset

We ran a 6-fold cross-validation experiment, where the 16 images were randomly split into 13 training images and 3 testing images. We compared segmentation accuracy achieved with our TST to the accuracy achieved without the TST. We used unweighted fixed thresholding (UFT) to fuse and threshold the images for the former, and weighted fixed thresholding (WFT-exp) and intensity thresholding (WIT-exp) for the latter. We also tested the effect of patch-based refinement when using a TST and intensity thresholding [13].

The use of a TST significantly improved segmentation accuracy from initially $77.94 \pm 1.77\%$ with UFT and no TST to $78.63 \pm 2.14\%$ with WFT-exp ($p = 1.5 \times 10^{-3}$) and further to $80.66 \pm 1.53\%$ with WIT-exp ($p = 2 \times 10^{-6}$) with a TST (Fig. 3.21). Patch-refinement improved accuracy for both fusion strategies ($p<.05$) with a significant effect between WFT-exp, which achieved $79.18 \pm 2.07\%$, and WIT-exp, which achieved $80.80 \pm 1.49\%$

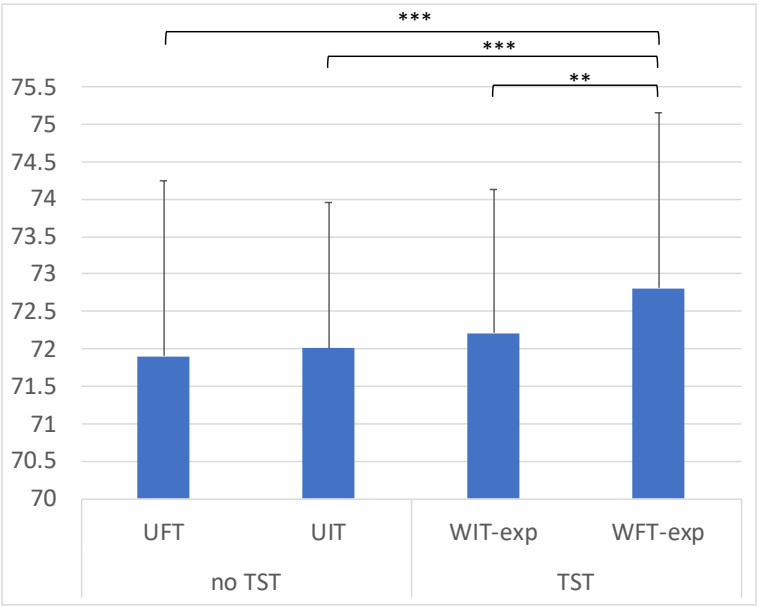


Figure 3.22: Improvement in segmentation accuracy by using our TST and different fusion and thresholding methods on the MICCAI 2012 dataset. Error bars represent the standard deviation.

($p = 1.4 \times 10^{-5}$).

3.5.8 Method validation on the MICCAI 2012 dataset

The dataset comprises 15 atlas and 20 target images and their corresponding label maps. We compared segmentation accuracy achieved with our TST to the accuracy achieved without the TST. We used UFT, WFT-exp, UIT and WIT-exp for candidate label fusion and thresholding for the former and UFT and UIT for the latter.

Our method reached a maximum Dice coefficient of 72.81 ± 2.33% over all labels and targets when using a TST and WFT-exp (Fig. 3.22). It achieved significantly better results compared to WIT-exp (72.21 ± 1.91%) with a TST ($p = 5.3 \times 10^{-3}$) and both unweighted variants, UFT (71.89 ± 2.35%) and UIT (72.00 ± 1.95%) when no TST was used (p<.001).

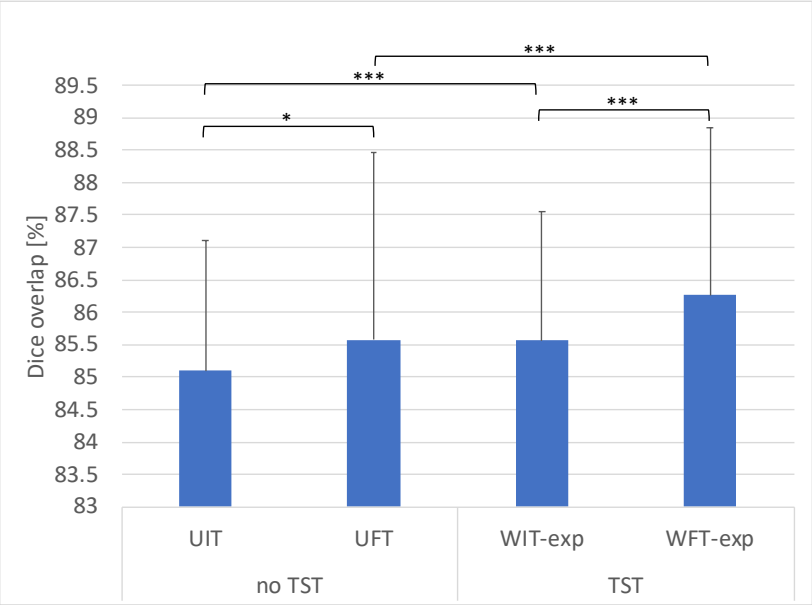


Figure 3.23: Improvement in segmentation accuracy by using a TST and different fusion strategies on the MICCAI 2013 dataset. Error bars represent the standard deviation.

3.5.9 Method validation on the MICCAI 2013 dataset

The dataset was divided into 3 subsets of 12 images each. In each fold of the cross-validation study one subset was used for testing with the remaining images as atlases. We compared segmentation results achieved without a TST to results with a TST. Unweighted fixed thresholding (UFT) and unweighted intensity thresholding (UIT) were used as fusion methods for the former. Both, weighted fixed tresholding (WFT-exp) and weighted intensity thresholding (WIT-exp) were used with exp and a TST.

The use of a TST and weights for the fusion significantly improved segmentation results compared to their unweighted variants without a TST ($p<.001$) (Fig. 3.23). The best result of $86.26 \pm 2.57\%$ was achieved with WFT-exp which showed significantly larger segmentation accuracy than intensity thresholding with ($p = 4.3 \times 10^{-5}$) and without the use of a TST ($p = 1.8 \times 10^{-2}$).

3.6 Discussion

In this section we tested linear and nonlinear dimensionality reduction methods for the reconstruction of a target image. The reconstruction of an atlas target with eigenimages showed high similarity to the original with only a small set of eigenimages. In contrast, a non-atlas target showed barely improvement, even with increasing number of eigenimages. This might be caused by the relatively small set of atlases for training and the use of linear methods for dimensionality reduction and registration. It could also be related to the global approach, where the eigenimages were calculated from the whole images, showing a large variability.

We improved our approach by using nonlinear registration methods, nonlinear dimensionality reduction methods, deformation fields as comparison basis and local ROIs for the construction of a TST with high similarity to the target. In our evaluation of nonlinear manifold embedding methods, Isomap showed slightly better results than LLE and was used for the remaining experiments. Duc *et al.* [75] achieved their best results with LLE but no significant difference to the results with Isomap were found. Although they presented a comprehensive evaluation of the parameters, their segmentation results were based on only one anatomical region. Parameters and overall segmentation results might differ between anatomical regions and regions of different size. In contrast we compared manifold embedding methods on 53 different anatomical ROIs, which provides a better overall performance evaluation. However, it still does not guarantee the selection of the best parameters for each ROI. In the future we are planning to investigate this challenging task further. Another difference to our study is their use of STAPLE [229] for label fusion (Section 4.3 for details), compared to weighted fixed thresholding and intensity thresholding in our approach. Since STAPLE has shown mixed results depending on the ROIs [11], its use with only one ROI might be more prone to introducing a bias towards the best parameter configuration. Another question that remains is how many atlases

are required to represent the manifold structure detailed enough to derive useful conclusions. In our experiments with Isomap we selected datasets with different cortical and non-cortical ROIs of varying size from healthy and diseased populations, which allowed tests with varying numbers of atlases. One drawback of our study is that the direct comparison between the eigenimage approach and nonlinear dimensionality reduction methods is difficult due to the use of different registration methods, and bases and scales of comparison. Although nonlinear methods, in general, outperform linear dimensionality reduction methods, it would be interesting to compare the results of eigenimages and Isomap side by side.

Consistent significant improvement in segmentation accuracy was observed with all datasets when our TST was used as an additional intermediate step for the propagation of the labels, which is due to two main reasons. Firstly, our TST is much more similar to the target than the average population template because of its regional construction process, and refines the registration between the target and population template. Consequently, the likelihood of registration errors is reduced. Secondly, the use of weights, determined from the distances in manifold space, and intensity thresholding allows more accurate fusion of the propagated labels. Intensity thresholding showed the best results for all datasets except for MICCAI 2012 and MICCAI 2013, where fixed thresholding was superior. This might be caused by reduced intensity contrast and poorly defined borders in sub-cortical ROIs, where the intensity profile is not as characteristic. It could also be related to the parameter selection for intensity thresholding. In Chapter 4 we will compare our overlap results to those from other methods as reported in the literature and discuss parameter selection for intensity thresholding in more detail.

Compared to other state-of-the-art MAS propagation and fusion strategies, our method is computationally more efficient. While other top performing methods require at least as many time-consuming nonlinear registrations as there are atlases, our method requires only two registrations at runtime.

In a runtime comparison for the MICCAI SATA challenge (15 atlases, 20 targets) our method required 2 nonlinear registrations with SPM's DARTEL per target at runtime. All other top 10 methods of the SATA challenge required pairwise registration of every atlas and target, i.e. 300 in total. In Chapter 4 we will also compare our results to methods from the literature. In our approach 3 deformation fields are composed, which are between an individual atlas and the population atlas, between the population atlas and the TST, and between the TST and the target. In the work by Sjöberg, Johansson and Ahnesjö [188], a decrease in segmentation accuracy was observed, linear to the number of composed deformation fields. However, in their work they used a population template to find correspondences between atlases and the given atlases as intermediate images. A deformable image registration method [124] with a coarse B-spline grid spacing of 8mm in all directions was used for the alignment. In our approach we improve the quality of composed deformation fields by using our TST with large local similarity to the target as intermediate image instead of random atlas images and by using a registration method with millions of degrees of freedom (3 x #voxels) instead of a coarse grid of control points.

Although, in general, the propagation of labels via composed deformation fields can be less accurate than via direct deformation fields, our results show that the selection and propagation of similar candidates with our TST as intermediate step can alleviate these drawbacks while increasing computational efficiency. In this chapter we have briefly introduced fusion methods, which will be discussed in more detail in the next chapter.

Chapter 4

Label Fusion

One of the essential elements of a MAS algorithm is the combination, i.e. fusion, of the propagated atlas labels (candidate labels). Based on the two approaches, single- and multi-atlas segmentation, described in Sections 1.2.2-1.2.3, the propagated atlas labels are either fused in average template space or in target space, respectively. In general, we can split combination strategies into global and local methods.

Global methods assign the same weight to the the labels of a candidate. One global method is majority voting, where the most frequently occurring label is assigned to a voxel. It has provided robust results in multiple studies [98, 99, 168]. Another popular method, initially developed for the fusion of manual segmentations and the evaluation of the human raters, is STAPLE [229]. In contrast to majority voting, weighted voting allows to assign a different weight to the labels of each candidate [12, 107]. The weight is usually calculated from the similarity between each candidate image and the target globally. Global combination strategies cannot guarantee that the selected candidate and its associated weights are the best choice for every ROI.

Locally weighted fusion is able to assign weights region-wise or voxel-wise. These include methods such as STEPS [47], joint label fusion [221] and patch-based label fusion [64]. However, the degree of locality is dependent upon

the local contrast of the ROI [11]. Voxel-wise weighting is highly sensitive to noise in low-contrast neighbouring ROIs, whereas global weighting is more robust. Conversely, in high-contrast ROIs local weights are superior.

4.1 Majority voting

The simplest label fusion strategy is majority voting which counts the votes for each of the L possible labels at every voxel x [11, 119]: $f_i(x) = \sum_{k=1}^{K} w_{k,i}(x)$ for $i = 1, ..., L$ and assigns the most frequently occurring label to it

$$E_{MV}(x) = max\,[f_1(x), ..., f_L(x)] \tag{4.1}$$

where

$$w_{k,i}(x) = \begin{cases} 1, & \text{if } i = e_k(x) \\ 0, & \text{otherwise} \end{cases} \tag{4.2}$$

Although very simple, it is a powerful and robust fusion method [169]. However, it does not incorporate a measure of similarity between target and atlas images.

4.2 Weighted voting

In contrast to majority voting, where each candidate's vote has the same power, weighted voting [12] allows to assign different weights to the candidates. Consequently, candidates, which are expected to yield a higher segmentation accuracy, get more power in the decision making. Label weights can be calculated locally or globally and are usually based on a chosen similarity metric. The global weights for expression 4.1 can be calculated as [11, 119]

$$w_{k,i}(x) = \begin{cases} m^p, & \text{if } i = e_k(x) \\ 0, & \text{otherwise} \end{cases} \tag{4.3}$$

where m is a measure of similarity and p a gain exponent to adjust the impact of the weight.

Local weights can be calculated as

$$w_{k,i}(x) = \begin{cases} [m(s,r)]^p, & \text{if } i = e_k(x) \\ 0, & \text{otherwise} \end{cases} \tag{4.4}$$

with the local similarity measure $m(s,r)$, where s is the shape of the patch and r is its radius.

4.3 Simultaneous truth and performance level estimation (STAPLE)

One of the main underlying problems of a MAS algorithm is the inter- and intra-operator variability introduced by manual segmentation, which is, however, still considered as the gold standard. Due to the propagation and fusion of these labels, one of the main error sources in MAS can be identified as inaccurate manual segmentation. STAPLE [229], initially designed to combine and rate expert manual segmentations and later applied in MAS, addresses this issue by computing a probabilistic estimate of the true segmentation and providing a measure of the performance level of each of the candidates. In the context of MAS a candidate is considered as an individual atlas. In line with the weighted voting algorithm, candidates with a higher performance should be assigned higher weights. However, the weights in STAPLE do not take the similarity between intensity images into account. Instead, the fusion algorithm calculates a confusion matrix. It uses an expectation-maximisation algorithm to calculate weights as a function of the estimated sensitivity p_i and specificity q_i characteristic for each candidate: $w_i = f(p_i, q_i)$ where the performance of a candidate can be summarised as $\boldsymbol{\theta_i} = (p_i, q_i)^T$. Here we consider the binary case, which can be extended to multiple labels. Optimal weights are determined based on the estimated performance level $\boldsymbol{\theta} = [\boldsymbol{\theta_1}, \boldsymbol{\theta_2}, ..., \boldsymbol{\theta_R}]$ of each of the R candidates, an a priori model for the spatial distribution of correlated ROIs and spatial homogeneity constraints

enforced by a Markov random field model. The goal is to maximise the log likelihood cost function of the complete data $(\mathbf{D}, \mathbf{T})$, where $\mathbf{T}$ is the true hidden segmentation and $\mathbf{D}$ holds the candidate segmentations and can be described as

$$(\hat{\boldsymbol{p}}, \hat{\boldsymbol{q}}) = \underset{p,q}{arg\,max}\, log\left(f(D, T|\mathbf{p}, \mathbf{q})\right) \tag{4.5}$$

The true segmentation $\mathbf{T}$ is unknown. The core algorithm iterates through two main expectation-maximisation (EM) steps. First, in the E-step, the conditional expectation of the true segmentation w_i^k to be 1 at voxel i and iteration k is estimated by

$$w_i^k \equiv f(T_i = 1|D_i, p^k, q^k) = \frac{a_i^k}{a_i^k + b_i^k} \tag{4.6}$$

with

$$a_i^k \equiv f(T_i = 1) \prod_j f(D_{ij}|T_i = 1, p_j^k, q_j^k) \tag{4.7}$$

$$b_i^k \equiv f(T_i = 0) \prod_j f(D_{ij}|T_i = 0, p_j^k, q_j^k) \tag{4.8}$$

Second, in the M-step, sensitivity and specificity (p, q) are estimated by

$$p_j^{k+1} = \frac{\sum_i w_i^k D_{ij}}{\sum_i w_i^k} \tag{4.9}$$

$$q_j^{k+1} = \frac{\sum_i (1 - w_i^k)(1 - D_{ij})}{\sum_i (1 - w_i^k)} \tag{4.10}$$

STAPLE was mainly designed to combine small numbers of candidate labels and has shown to produce mixed results depending on the ROIs and dataset, for which it received criticism [11, 66]. Another limitation of STAPLE is its purely probabilistic approach, which does not take local intensity similarity measurements into account. In an improved version of the method called non-local STAPLE [19], the initial statistical fusion algorithm was merged with a non-local approach that incorporates patch-wise weight estimation (for more details on patch-based fusion see Section 4.8).

4.4 Selective and iterative method for performance level estimation (SIMPLE)

SIMPLE [129] aims to combine the advantages of atlas selection, where only a subset of the most promising candidates is used for fusion, and the truth and performance level estimation presented by STAPLE. SIMPLE's core algorithm, just like STAPLE, iterates through estimation of performance of each rater and estimation of the ground truth segmentation, but, in contrast to STAPLE, candidates with a low performance are not considered in future iterations. The exclusion of bad candidates limits their influence on the final segmentation. The algorithm starts by estimating the performance of each candidate image as its NMI similarity to the target image. Then all candidate labels are fused into a ground truth estimate L_{est} with majority voting. These initial performance and ground truth estimates are refined by iterating through 3 main steps: First, the performance of each candidate label at iteration k is estimated by measuring the overlap of each candidate label L_i and L_{est} as $\Phi_i = f(L_i^{'}, L_{est}^{k})$. Second, candidate labels are only included in further iterations $L_i^{'} \epsilon \tilde{L}^{k+1}$ if $\Phi_i > \Theta$. Third, the new segmentation L_{est}^{k+1} is calculated with majority voting.

SIMPLE has shown to outperform STAPLE and locally weighted methods based on NMI, but two main drawbacks can be identified. Due to its strong dependence on the initial ground truth estimate, the method could fail if the candidate labels are poorly aligned. And the method works globally, which might lead to the exclusion of useful local information.

4.5 Similarity and truth estimation for propagated segmentations (STEPS)

Another method which evolved from STAPLE is STEPS [47]. In contrast to the global approach in STAPLE, STEPS selects and fuses the locally

best ranked templates based on locally normalised cross correlation (LNCC), for which a new hidden parameter was introduced, which equals 1 if the candidate is top ranked for a particular location. The concept of sensitivity and specificity was extended for the use of multiple classes, by introducing a confusion matrix which provides a measure of agreement or disagreement between the candidate segmentations and the fused segmentations.

4.6 MALP

A different approach was taken by Ledig *et al.* [131], in which the spatial information provided by MAS is only used as spatial prior for an intensity model solved by EM optimisation. Assuming that the voxel intensities belonging to one of the K labels are normally distributed, the goal can be formulated as finding the unknown segmentation z_i based on the image intensity distribution. The conditional probability of finding an intensity y_i at a voxel i can be computed as

$$f(y_i|\boldsymbol{\phi}) = \sum_k f(y_i|\mathbf{z_i} = \mathbf{e_k}, \boldsymbol{\phi}) f(\mathbf{z_i} = \mathbf{e_k}) \tag{4.11}$$

where $f(\mathbf{z_i} = \mathbf{e_k})$ is the probability of voxel i belonging to label k and $f(y_i|\mathbf{z_i} = \mathbf{e_k}, \boldsymbol{\phi})$ is the Gaussian distribution of intensity values belonging to label k described by mean and standard deviation $\boldsymbol{\phi} = \{(\mu_1, \sigma_1), ..., (\mu_K, \sigma_K)\}$. The EM algorithm iterates through calculating the label probability and maximising $f(y|\phi)$. The model was further refined by incorporating a priori spatial information from MAS using Markov random fields (MRF).

4.7 Joint label fusion

Most weighted atlas fusion strategies calculate the weight for each candidate label independently as the similarity between the candidate image and the target. Consequently, the same label errors, produced by different candidates, are difficult to detect and will contribute to the final segmentation.

The aim of joint label fusion [221] is to reduce these correlated segmentation errors by finding the optimal weights to reduce the expected total error. The joint error distribution of two candidate labels is estimated based on the similarity between their respective candidate images and the target. The voting weights for n candidates are obtained by

$$\mathbf{w}_x = \frac{M_x^{-1} 1_n}{1_n^t M_x^{-1} 1_n} \tag{4.12}$$

where $1_n = [1; 1; ...; 1]$ and M_x is the pairwise dependency matrix, which is estimated from the intensity similarities of each candidate pair (F_i, F_j) and the target image F_T:

$$\begin{aligned} M_x(i,j) &= p(\delta^i(x)\delta^j(x) = 1 | F_T(y), F_i(y), F_j(y) | y \epsilon N(x)) \\ &\propto \left[\sum_{y \epsilon N(x)} |F_T(y) - F_i(y)||F_T(y) - F_j(y)| \right]^{\beta} \end{aligned} \tag{4.13}$$

where $N(x)$ is the neighbourhood for comparison around a voxel x.

The approach was further extended [222] by a method to correct systematic errors [220], introduced by the segmentation model or optimisation algorithm. In the case of one label, one AdaBoost classifier is trained on the atlas image set to discriminate between correctly and incorrectly labeled voxels. After the candidates are fused, the classifier checks the label of each voxel and potentially corrects it.

4.8 Patch-based label fusion

The joint label fusion algorithm requires the estimation of a dependency matrix, which is created by measuring the similarity of the local neighbourhood around each voxel in the candidates and the target. This simple local similarity measurement method is generally referred to as patch-based comparison. The patch-based algorithm is based on Buades *et al.'s* [41] non-local means

algorithm, initially designed for image denoising, and has later been adapted for the use as label fusion method [64]. The method compares the patch $P(x_i)$ around each voxel x_i in the target image to the patches $P(x_{s,j})$ within a search volume V_i around the corresponding voxel $x_{s,j}$ in each of the N atlas images. Based on image intensity similarities between patches, weights $w(x_i, x_{s,j})$ are assigned to the corresponding atlas labels $y_{s,j}$ resulting in the probability

$$v(x_i) = \frac{\sum_{s=1}^{N} \sum_{j \epsilon V_i} w(x_i, x_{s,j}) y_{s,j}}{\sum_{s=1}^{N} \sum_{j \epsilon V_i} w(x_i, x_{s,j})} \tag{4.14}$$

with the weights

$$w(x_i, x_{s,j}) = \begin{cases} exp^{\frac{-\left\|P(x_i)-P(x_{s,j})\right\|_2^2}{h}} & \text{if} \quad ss > th \\ 0 & \text{else} \end{cases} \tag{4.15}$$

Atlas patches are pre-selected based on the structural similarity $ss = \frac{2\mu_i\mu_{s,j}}{\mu_i^2+\mu_{s,j}^2} \times \frac{2\sigma_i\sigma_{s,j}}{\sigma_i^2+\sigma_{s,j}^2}$ [227] to the target patch to discard potentially unsuitable patches when it is less than a pre-defined threshold th. Additionally, a pre-selection step can be applied to discard atlases as a whole and the ROI for patch-comparison can be reduced to the specific anatomy to improve efficiency. A main advantage of the method is the use of redundant information from an increased number of samples, taken from different atlases, to make the decision more robust. Consequently, the impact of registration errors is reduced due to the patch comparisons within a local neighbourhood. The method was initially proposed with only linearly aligned atlases without requiring a time-consuming nonlinear registration.

A more efficient approach has been proposed by Rousseau *et al.* [174], where, instead of comparing patches around every voxel, only a subset of patches is considered and, rather than determining a label only for the centre voxel of each patch, a label estimate can be provided for every voxel within the patch. The basic patch-based approach has been further extended by modelling a

patch as a sparse linear superposition of patches [237, 243], which can be solved with linear regression. The resulting reconstruction coefficients can then be used as weights for the label fusion. A probabilistic patch-based model was proposed by Bai *et al.* [25], which incorporated label information into the registration process to improve both segmentation and registration accuracy.

4.9 Our approach to propagation, fusion and binarisation

Equipped with a TST, we can now non-linearly register the affinely transformed target image $\widetilde{T}$ to it to get deformation field $D_{\widetilde{T}\to\mathrm{TST}}$ using SPM's DARTEL in a non-iterative fashion, for maximum efficiency. Since the TST is very similar to $\widetilde{T}$, this non-linear registration should be very accurate.

For each label r we then warp all individual label maps $\{\widetilde{L_j^r}\}_j$ onto the target by using the composed transformation $D_{\widetilde{T}\to\mathrm{TST}} \circ D_{\mathrm{TST}\to\bar{I}} \circ {D_{\widetilde{I}_j\to\bar{I}}}^{-1}$ and trilinear interpolation to get N candidate segmentations $\{\widehat{L_j^r}\}_{j=1\ldots N}$.

We then use a local fusion strategy and compute for each label the (1) unweighted (UW), where each candidate segmentation receives the same weight, and (2) weighted (W) sum of the candidate segmentations with the same weights we used for TST construction: $\widehat{L^r} = \sum_{j=1}^{N} \omega_j^r \cdot \widehat{L_j^r}$ with $\sum_{j=1}^{N} w_j^r = 1$. Note that the unweighted method uses the probabilistic values for fusion and ranked candidates, compared to majority voting, which uses binarised labels and counts the occurrence. The weights for the weighted scheme were calculated either with $w_j^r = \frac{1}{D_j}$ (1/D) or $w_j^r = e^{\frac{-D_j}{std(D)}}$ (exp) for comparison, where D is the distance matrix and D_j is the distance between the target instance and the j-th candidate in manifold space. We project those back onto the space of the input target image, by applying the inverse of the previously estimated affine transformation using trilinear interpolation.

Finally, we threshold the probabilistic values so they can be compared

against ground truth delineations. Four thresholding strategies were implemented and tested:

1. A fixed threshold (FT) θ is used to binarise the probabilistic values:

$$Y(x) = \begin{cases} 1 & if \quad \widehat{L^r}(x) > \theta \\ 0 & else \end{cases} \tag{4.16}$$

2. We use the probabilistic values and also consider the intensity values of the target to further refine the segmentation (IT). We check, similarly to Doshi *et al.'s* approach [74], whether the intensity of a voxel falls within the intensity distribution, given by the mean and a multiple c of the standard deviation, and estimated from those areas of the target image with a probability of at least $\delta\%$ of the maximum probability: $Y = \{x|\widehat{L^r}(x) > \delta \cdot max(\widehat{L^r})\}$. Empirical tests showed that this refinement did not systematically improve performance for non-cortical regions, which is consistent with Ledig *et al.*'s findings [131] and is probably due to the fact that tissue contrast around cortical regions is much stronger than around subcortical structures. We further improved the method by checking if the probability of the identified voxels exceeds the threshold defined by $\epsilon_r(x) = std(\widehat{L^r})$.

3. A local threshold (LT) $\delta_r(x) = \text{std}_{y \in N(x)}(\widehat{L^r}(y))$ is calculated from the probabilistic label maps where $N(x)$ is a neighbourhood around voxel x whose size is dictated by that of the label, with larger labels yielding larger neighbourhoods (see Section 1.2.1.1).

4. In [13] we presented an efficient hierarchical method based on our approach with a TST. After regional weighted candidate fusion with a fixed threshold we further refine regions below a predefined threshold, since these are usually more likely to be misclassified. In these regions we recalculate the weights locally with a patch-based approach (Section 4.8), providing a second set of locally refined classification probabilities.

4.10 Experiments

We tested several fusion methods with the approach described in Section 3.4. For some of them we provide the segmentation accuracy achieved without a TST as a reference. The used parameter values for the fusion and thresholding methods will be provided for each experiment.

4.10.1 Evaluation of fusion and thresholding strategies on the LONI dataset

Various fusion strategies were tested in a cross-validation experiment with four subsets of 30 atlas and 10 target images each. One subset was used to empirically determine the parameters for intensity thresholding by varying the values for the coefficient c and the threshold δ. The intensity distribution was sampled from those areas in the target with the δ highest likelihood of being part of the label, as given by the fused candidate labels. Then two variants of the method were tested. First, each target voxel within the search region, defined by probability values larger than zero, was compared to the distribution to make a final decision. Second, the search region was further limited to voxels with a probability of at least the standard deviation of the fused label probabilities. Once the best parameters were found, the fusion and thresholding strategies MV, WFT-exp, WFT-1/sim, WIT-exp and WIT-1/sim were compared with a fixed threshold $\theta = 0.33$ for FT methods, and $\delta = 0.4$ and $c = 4$ for IT methods.

Our variant of using a refined search region (wt) showed improvement over the basic variant of using just the intensity distribution (Fig. 4.1). A steep incline can be noticed from $c = 1$ to $c = 2$ for both variants and all choices of δ, which flattens faster for higher values of δ. The basic variant showed a decline with increasing coefficients, which is not the case with the refined variant, which is also governed by the label probability. Consequently, for larger values of c and δ the intensity distribution has less impact. The best result was obtained with c=4 and θ=0.4. In contrast to other datasets,

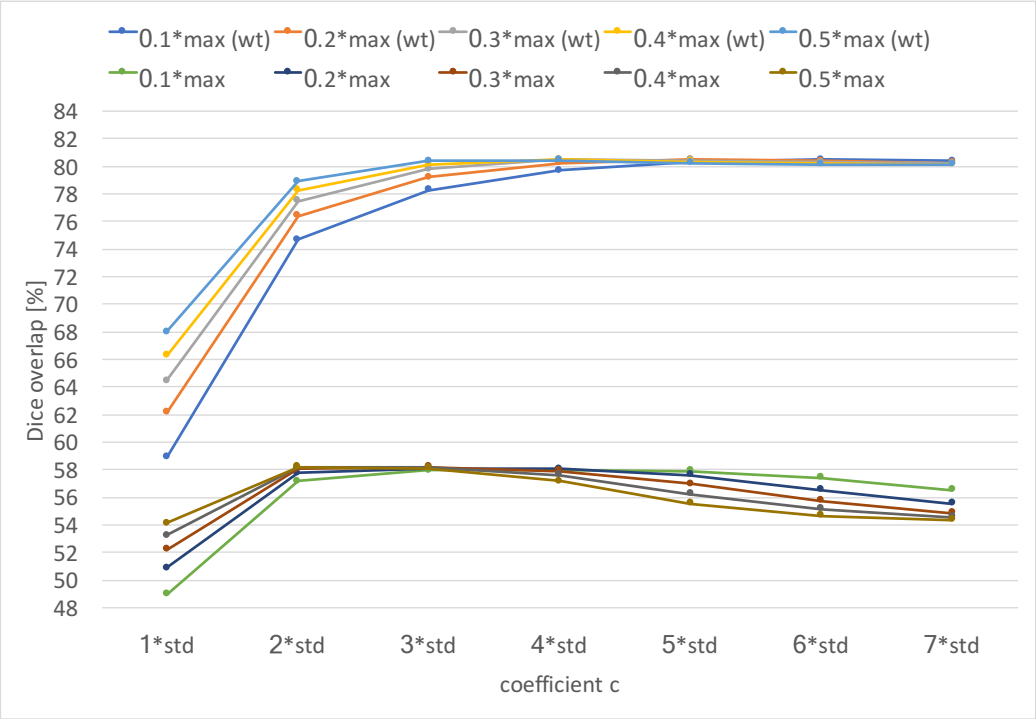


Figure 4.1: Empirical evaluation of the intensity thresholding parameters. The coefficient c is systematically increased on the x-axis with varying values for δ. Both intensity thresholding variants, with (wt) and without refinement of the search region, were tested.

for which we consistently used c=3 and δ=0.1, these parameters have large values which is due to the wide intensity distribution of the LONI labels. Some of these labels contain multiple tissue types such as GM and WM.

A segmentation accuracy of 79.14 $\pm$1.24% was achieved with MV without a TST and is provided as a reference (Fig. 4.2). The use of a TST did not improve accuracy when using MV. It might not be sensitive enough to detect changes in the probability due to the binarisation of the probabilistic values before the label frequency count is performed. However, weighted methods such as WFT and WIT, which use the probabilistic values, were capable of detecting these changes. Significant improvements in segmentation accuracy were achieved with WFT-exp over MV ($p = 4.8 \times 10^{-7}$), WFT-1/sim over WFT-exp ($p = 9.8 \times 10^{-10}$), WIT-exp over WIT-1/sim

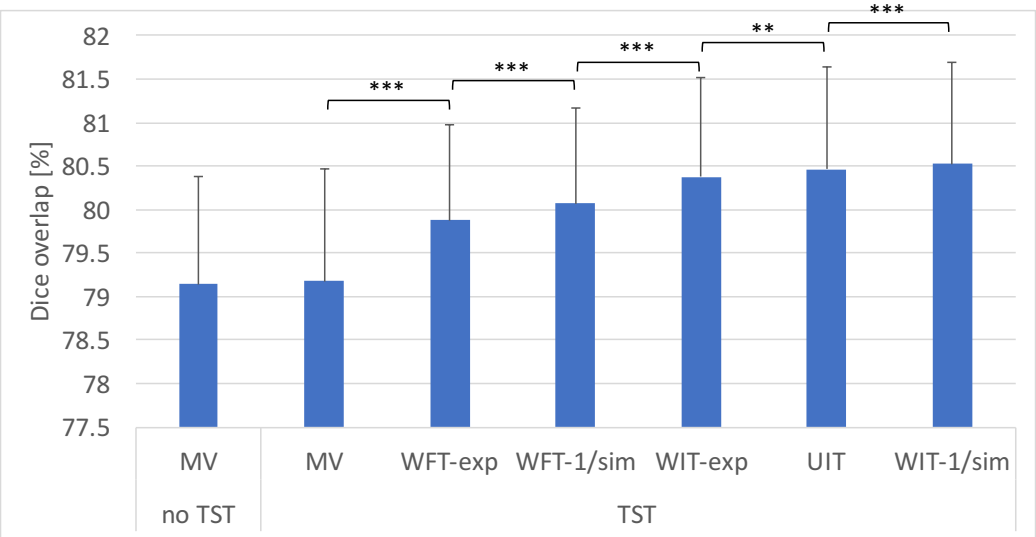


Figure 4.2: Comparison of fusion and weighting strategies on the LONI dataset with a TST. The overlap achieved with MV without a TST is given as a reference. Error bars represent the standard deviation.

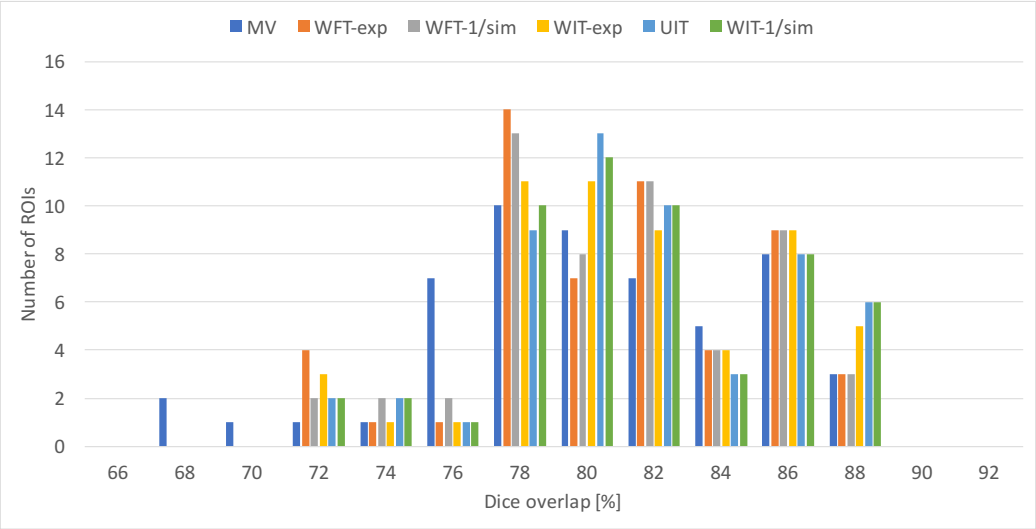


Figure 4.3: The histogram outlines the number of anatomical ROIs of the LONI dataset that fall within a certain range of accuracy. For each bin and for each fusion method the number of ROIs is provided.

($p = 3.9 \times 10^{-7}$), UIT over WIT-exp ($p = 9.4 \times 10^{-3}$) and WIT-1/sim over UIT ($p = 4.0 \times 10^{-7}$). The best result of $80.53 \pm 1.17\%$ was achieved with

WIT-1/sim weighted voting and intensity thresholding. Although a statistically significant improvement was found between the three weighted and unweighted intensity thresholding methods, UIT showed with 80.47 ± 1.18% a high level of accuracy. Since intensity thresholding also takes the intensity distribution into account, it can alleviate the impact of registration errors. The cortical labels of the LONI dataset contain white and grey matter tissue which might not allow a clear classification of voxels due to the large intensity distribution in high-probability ROIs. Another reason that might explain the inferior result of WIT-exp compared to UIT is the quality of the registrations. In Klein *et al.'s* [122] evaluation of registration methods DARTEL was ranked sixth for the LONI dataset, while performing excellent on two other datasets. Consequently the deformation fields might not be precise enough to derive reliable weights with WIT-exp.

The histogram in Fig. 4.3 shows the accuracy improvement for the LONI ROIs by using fixed and intensity thresholding methods. For MV and fixed thresholding methods a large proportion of the 53 ROIs reached less than or equal 78% accuracy. With the use of intensity based methods a shift in the number of ROIs that reached a higher accuracy is noticeable, where WIT-1/sim and UIT have the highest number of ROIs above 86% compared to other fusion methods. Noticeable is the large number of ROIs for MV at the 84-86% bins, which count more or at least an equal number of ROIs compared to the highest ranked methods.

This compares favourably against the results reported by Wu *et al.* [237] who implemented three patch-based methods, their own (SCPBL) as well as that by Coupe *et al.* [64] (PBL) and by Rousseau *et al.* [174] (SPBL), and tested all three on the LONI dataset. The reported Dice scores were 75.06 ± 2.35% for PBL, 76.46 ± 1.96% for SPBL and 78.04 ± 1.34% for SCPBL, all inferior to ours. This also outperforms the recently proposed method by Zikic *et al.* [245], who reached 80.14 ± 4.53%, for a comparable computation time. Note, even our simpler weighted voting method with fixed thresholding achieved comparable results of 80.07% ± 1.1% to Zikic's method. Our

method was slightly inferior to the one presented by Wu *et al.* [236], which achieved 81.46 ± 2.25% but used only one set of test images for evaluation while we performed cross-validation experiments. Better results were also achieved by Ma *et al.* [136] with 82.56% ± 4.22% at great computational cost, since they require pairwise registration between each atlas and target, and the comparison of local patches. Considering one subset of 30 atlases and 10 targets it would take 300 nonlinear registrations at runtime with their method, compared to only 20 nonlinear registrations with our method.

4.10.2 Evaluation of fusion and thresholding strategies on the ADNI-HarP dataset

The dataset was divided into 5 subsets of 26 images each and a TST was constructed for each target. Four methods (UFT, WFT-exp, UIT, WIT-exp) were tested with 1, 3, 5, 10, 15 and 40 candidates on one of the subsets. The fixed threshold θ was set to 0.35 for the methods with FT. For methods with IT, δ was set to 0.1 and the coefficient c to 3. With the same settings, we performed a cross-validation experiment on all subsets for 5, 10, 15 and 40 candidates. It should be noted that even though each candidate was assigned the same weight in the unweighted methods, the ranking and selection of the candidates was performed with manifold learning.

The overall best result of 84.64 ± 2.89% for both ROIs on one subset was achieved with 15 candidates and unweighted intensity thresholding (UIT). Both the unweighted (UIT) and weighted intensity thresholding (WIT-exp) methods showed a significant improvement in segmentation accuracy over unweighted (UFT) and weighted fixed thresholding (WFT-exp) for all candidate choices ($p<.001$). The use of all 40 candidates lead to a slight decrease in accuracy for both unweighted methods but an increase for WIT-exp. The observed decrease is consistent with the findings in the literature and can be explained by the introduction of unsuitable atlases and, in turn, segmentation errors. The increase might be caused by the weighting scheme, which is

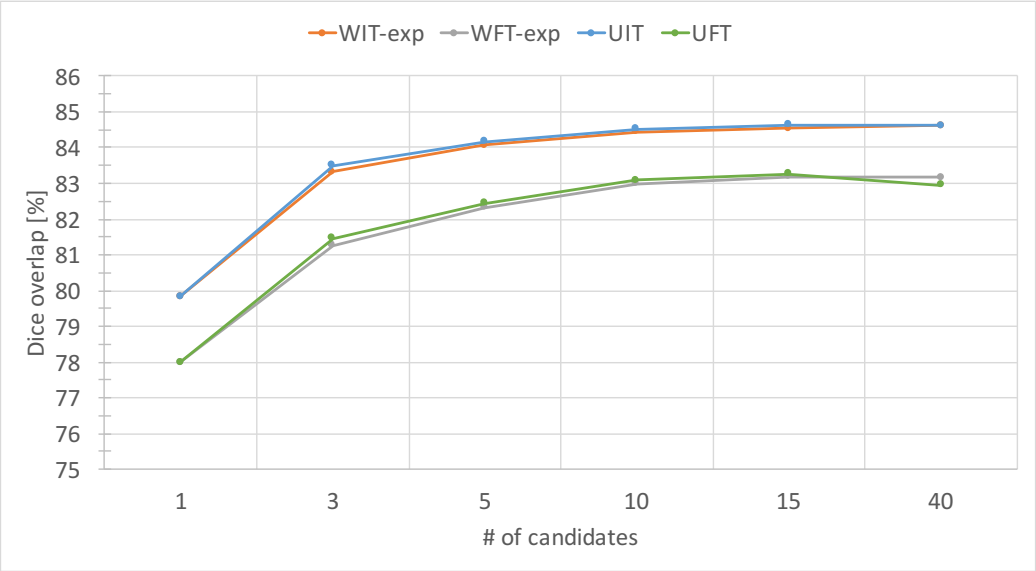
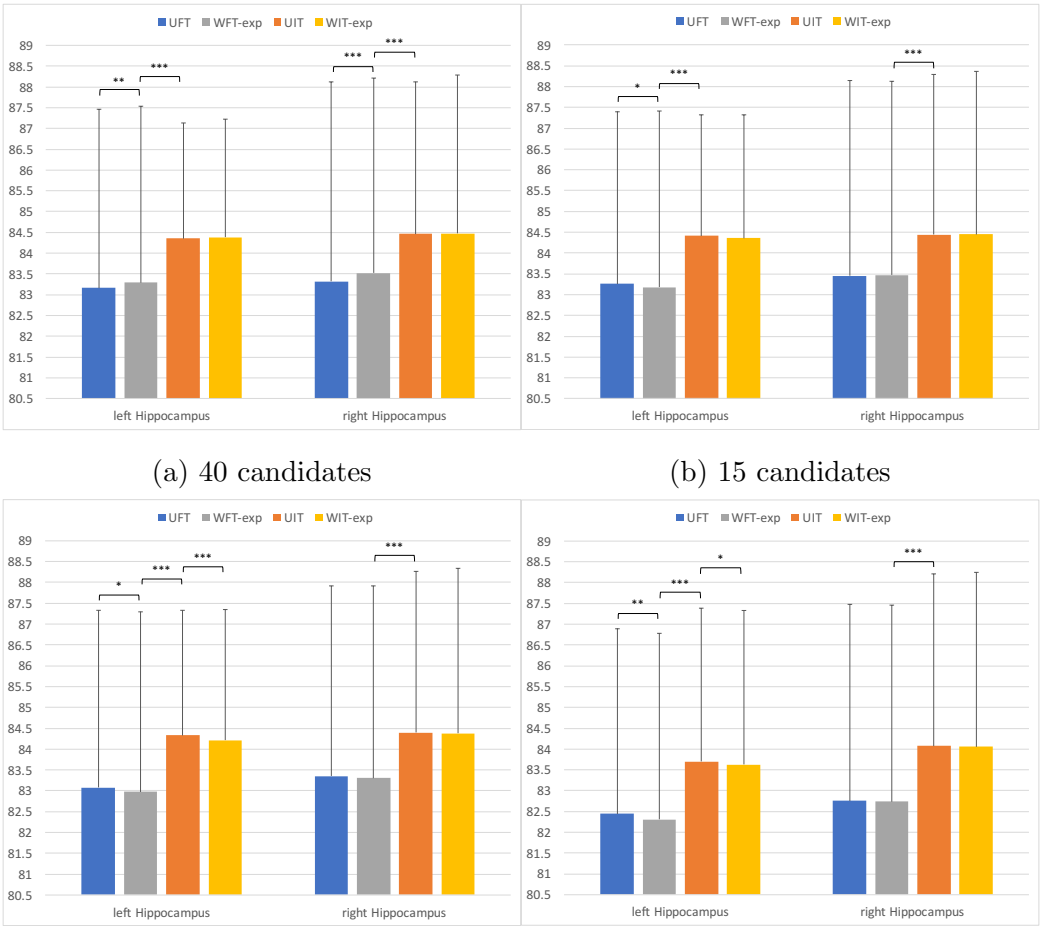


Figure 4.4: Comparison of fusion strategies and weighting schemes with varying numbers of candidates on one subset of the ADNI dataset with a TST.

capable of identifying unsuitable atlases and assigning them a lower weight. When using only 5 candidates and UIT, we still reached a Dice overlap over 84%. When further reducing the number to 3 candidates, intensity-based methods were still superior (83.5%) compared to the best overall results achieved with fixed thresholding (83.27%). This shows that intensity thresholding is efficient, robust to the number of candidates and to low intensity contrast ROI such as the hippocampus.

Surprisingly, the unweighted strategies (UFT, UIT) achieved similar or even slightly better results than the weighted strategies (WFT-exp, WIT-1/exp) for 15, 10 and 5 candidates. One reason might be that empirical tests for determining the intensity thresholding parameters were performed with 40 candidates, where the weighted strategies slightly outperformed the unweighted strategies. Another reason for the similar results achieved with unweighted and weighted methods might be that the ranking and selection of the most similar atlases in manifold space was performed for both strategies. The large number of atlases in the training set increases the chance of finding similar candidates to the target, which would lead to evenly distributed



(a) 40 candidates

(b) 15 candidates

(c) 10 candidates

(d) 5 candidates

Figure 4.5: Evaluation of fusion methods for each ROI and different candidate numbers for the ADNI dataset. Error bars represent the standard deviation.

weights for the weighted strategies. Consequently, the use of weighted methods would have a similar effect as the use of unweighted methods, where all candidates are assigned the same weight.

The ADNI-HarP dataset was also used in Platero and Tobar's study [157] for the evaluation of their registration- and patch-based method. They presented the results achieved with Freesurfer [80] and their own approach. Freesurfer reached an overall Dice coefficient of $78.2 \pm 4.1\%$, which is inferior to our maximum of $84.64 \pm 2.89\%$. In order to obtain a similar level of accuracy as Freesurfer, our method would require only 1 candidate with a fixed thresholding method. However, it should be noted that Freesurfer uses a training dataset with different anatomical definitions. Platero's own method achieved $85.00 \pm 4.5\%$ which is slightly better than our best results. However, it requires the computationally expensive direct nonlinear registration between each selected atlas and the target and additionally patch-refinement. Patch-based refinement as a post-processing step could also be added to our processing pipeline as outlined in Section 4.9 and tested in Section 4.10.4. Benkarim *et al.* [31, 30] used a set of ADNI images of similar size to ours and tested MV, STAPLE, STEPS, joint label fusion and their own learning-based method SCMWF2 with different registration methods. In the best configuration they reached $76.7 \pm 4.9\%$, $76.8 \pm 5.8\%$, $79.9 \pm 4.3\%$, $86.0 \pm 3.7\%$ and $86.6 \pm 2.6\%$ respectively. Our results are superior to all except joint label fusion and SCMWF2. However, in SCMWF2 the correspondence between each of the atlases and targets was achieved via the creation of an average population atlas and the composition of the corresponding deformation fields. This population atlas was constructed from both atlas images and target images. Consequently, to find the correspondence between a new target and each atlas, an average population template would have to be constructed for each target at runtime, leading to enormous computational costs. The only alternative would be to directly register each atlas and target at runtime. In both articles SCMWF2 was only tested on datasets with sub-cortical ROIs.

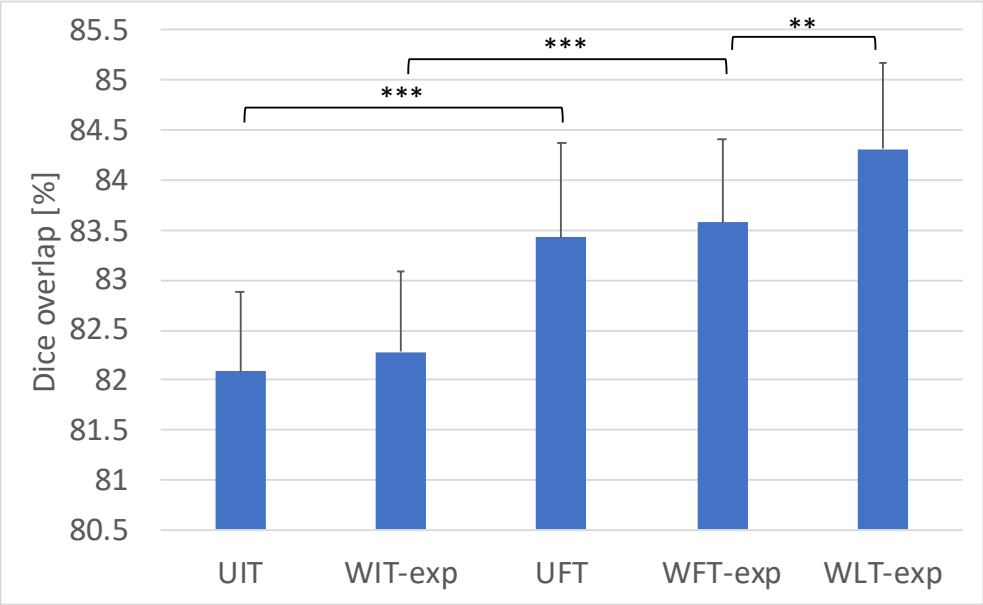


Figure 4.6: Comparison of fusion and weighting strategies on one subset of the IBSR dataset with a TST. Error bars represent the standard deviation.

4.10.3 Evaluation of fusion and thresholding strategies on the IBSR dataset

The IBSR dataset was divided into 6 subsets of 3 images each. UIT, UFT, WIT, WFT and weighted local thresholding (WLT) was applied to all subsets after constructing a TST for each target. The fixed threshold θ was set to 0.35 and the intensity thresholding parameters to $\delta = 0.1$ and $c = 3$. The CSF label was excluded as it has been shown to introduce a bias [215].

Both weighted methods (WFT-exp, WIT-exp) outperformed their unweighted variants (UFT, UIT) (Fig. 4.6) but no statistcally significant difference was found ($p>.05$). Both fixed thresholding methods showed significantly higher segmentation accuracy than the intensity thresholding method ($p<.001$). This could be caused by the dataset-specific labels, which contain multiple tissue types within the same label, e.g. the GM label also contains parts of CSF, WM and background, and poor contrast between ROIs. This also questions the viability of the dataset for the measurement of segmenta-

tion accuracy. The best results of 84.31 ± 1.35% were obtained with weighted local thresholding (WLT), which outperforms Zikic *et al.'s* method [245], which achieved 83.5 ± 4.2%. We also outperformed the method by Rousseau *et al.* [174], which reached an overall Dice score of 83.5%. The performance of our method is however slightly inferior to that recently reported by Doshi *et al.* [74], ranked first in the MICCAI 2013 segmentation challenge, and which reached 84.96 ± 1.3% with similarity ranking and boundary modulation. However, their approach requires two registrations between each of the selected atlas images and the target at runtime, which makes it computationally very expensive. Considering one target image and 15 atlases, it would require them 30 nonlinear registrations at runtime, compared to only 2 nonlinear registrations with our method. They also reported their results when using MV and global similarity ranking without the boundary modulation, for which they achieved 83.23 % ± 1.36 % and 84.14 % ± 1.3 %, respectively. Both are inferior to our results at much higher computational costs.

4.10.4 Evaluation of fusion and thresholding strategies on the NIREP-NA0 dataset

The dataset was divided into 6 subsets of 3 images each. In a cross-validation experiment a TST was constructed for each target and two thresholding methods (WFT-exp, WIT-exp) were tested with all candidates. The fixed threshold θ was set to 0.35 for FT methods. For methods using IT, δ was set to 0.1 and the coefficient c to 3. In addition, we tested a larger δ of 0.26 and evaluated the impact of patch-refinement for regions with a probability below 0.4.

Figure 4.7 shows a slice through a target scan with the manual ground truth label in yellow (left) and the fused candidate segmentations in green (middle). Based on the pre-defined threshold we divided the fused label into high and

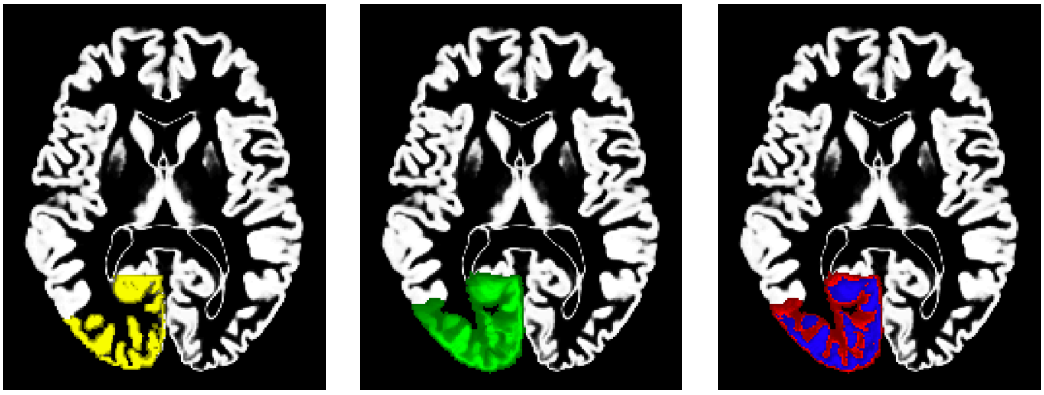


Figure 4.7: Slice through one of the target images with one ground truth label (left), the fused candidate segmentations (middle), and high (blue) and low (red) probability regions derived thereof (right).

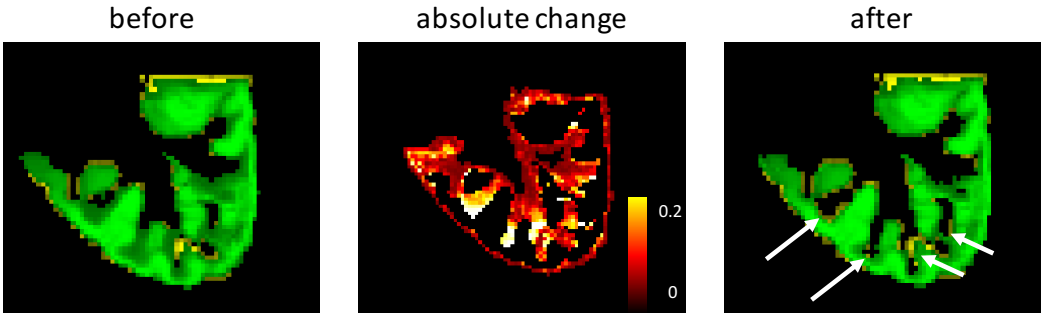


Figure 4.8: Enlarged ROI with the probabilistic label created with weighted fusion of the candidate segmentations before the patch-refinement (left), the absolute change in probabilities achieved by the patch-refinement (middle) and the final segmentation result after the patch-refinement (right). The white arrows indicate regions where substantial improvement was achieved.

low probability regions as indicated in blue and red respectively (right). For this particular ROI the number of patch-comparisons could be reduced by a half from approximately 96000 to 48000.

The improvement by using the patch-refinement is illustrated in Figure 4.8. After the refinement the label provided a more detailed and crisper outline of the anatomical structure.

A significant improvement ($p<.001$) was observed between weighted fixed

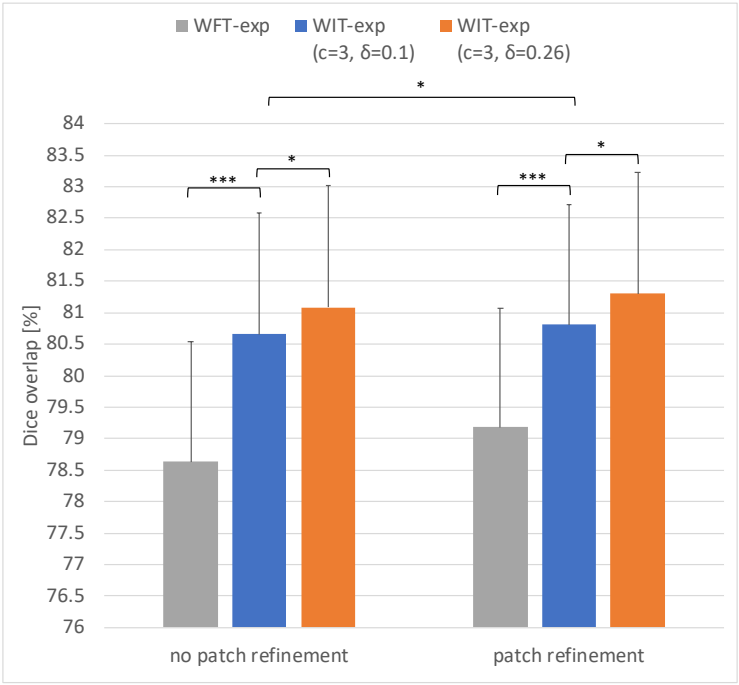


Figure 4.9: Comparison of fusion and weighting strategies on all subsets of the NIREP dataset with a TST. Error bars represent the standard deviation.

thresholding (WFT-exp) and weighted intensity thresholding (WIT-exp, δ=0.1) with patch-refinement and without patch-refinement (Fig. 4.9). Increasing δ to 0.26 further significantly improved results ($p<.05$). The use of additional patch-refinement had a significant effect on segmentation accuracy achieved with all WFT-exp and WIT-exp ($p<.05$) methods. Each bin of the histogram in Fig. 4.10 shows the number of ROIs that reached its corresponding Dice overlap. A large proportion of the ROIs reached less than or equal 78% with WFT-exp with and without patch-refinement. The use of WIT-exp with patch-refinement shows a shift in the number of ROIs towards higher accuracy bins. The positive impact of patch-refinement was mainly observed in regions with low segmentation accuracy ($\sim$ 76-78%), which achieved medium segmentation accuracy ($\sim$ 80%) after the refinement.

Our best result of 81.3% $\pm$ 1.91% was achieved with a TST, weighted intensity thresholding (WIT-exp, δ=0.26) and patch-refinement and took approximately 1.5h per image with 32 ROIs. This compares very favourably

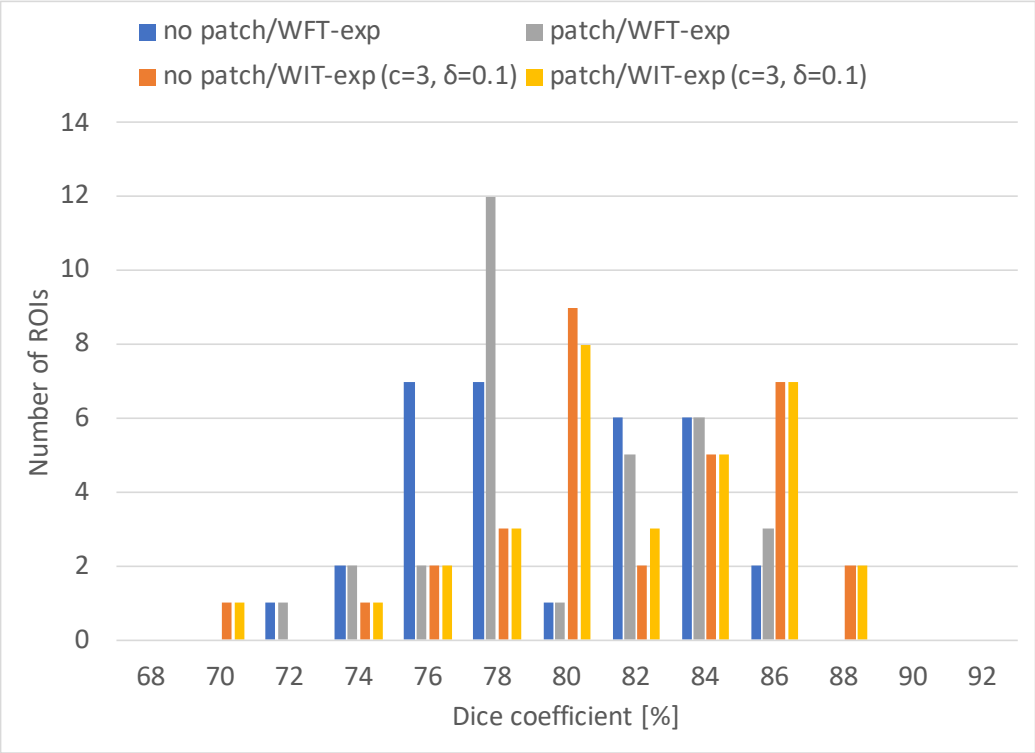


Figure 4.10: The histogram outlines the number of anatomical ROIs of the NIREP dataset that fall within a certain range of accuracy.

against the registration-based method by Doshi *et al.* [74], the top-performer on the MICCAI 2013 challenge. Their best result of 80.95 ± 1.79% was achieved with 7 selected atlases and required 14 pairwise non-linear registrations per target, 7 with each of the two registration methods (ANTS, DRAMMS). In contrast, our method required only two non-linear registrations per target. We also outperform the recently published patch-based iterative approach by Wang *et al.* [226] who reached 77.45 ± 3.39%. Wu *et al.* [237], who implemented three patch-based methods, their own as well as that by Coupe *et al.* [64] and by Rousseau *et al.* [174] reported Dice scores of 73.37 ± 3.25%, 74.58 ± 2.87% and 76.33 ± 2.25% respectively, all inferior to ours. For a typical ROI, runtimes for those methods were 10 min, 28 min and 45 min respectively, resulting in an approximate minimum runtime of 5.3h and maximum of 24h for all ROIs, which is in strong contrast to our 1.5h.

The sparse patch-based method by Zhang *et al.* [243] achieved 75.6% and the recently proposed forward and backward patch-based method by Sun *et al.* [199] reached 77.06%, both inferior to ours. The performance of our method is however slightly inferior to the results by Wu [238] with 82.7%, who used pairwise deformable registrations and joint probabilities, but required 7.7h per target. Similarly, with the use of ANTs and STAPLE in [174], which are both very time-consuming, a Dice overlap of 82.3% was reached.

4.10.5 Evaluation of fusion and thresholding strategies on the MICCAI 2012 dataset

The dataset comprises 15 atlases and 20 target images. In a cross-validation experiment a TST was constructed for each target and four combinations of weighting and thresholding methods (UFT, WFT-exp, UIT, WIT-exp) were tested with all candidates. The fixed threshold θ was set to 0.35 for the methods using FT. For IT methods we kept c at 3 and tested 0.1 and 0.17 for δ.

A significant improvement in segmentation accuracy was observed between UIT and WIT-exp for both parameter configurations with δ=0.1 ($p = 3.2 \times 10^{-3}$) and δ=0.17 ($p = 0.025$) (Fig. 4.11). The fixed thresholding methods, UFT and WFT-exp, achieved a significantly larger Dice overlap than intensity thresholded methods when δ=0.1 was used ($p = 5.3 \times 10^{-3}$), but a significantly smaller Dice overlap with δ=0.17 ($p = 1.04 \times 10^{-3}$). The best result of 73.51% $\pm$ 1.99% (82.19 $\pm$ 2.46% for non-cortical and 70.32 $\pm$ 2.00% for cortical) was achieved with WIT-exp and δ=0.17.

Each bin of the histogram in Fig. 4.12 shows the number of ROIs that reached its corresponding Dice overlap. The largest shift from lower towards higher Dice overlap bins can be observed between 70% and 80%. Intensity thresholded methods with δ=0.1 dominate lower accuracy bins around 55%. Increasing δ to 0.17 shows a larger number of ROIs with a high segmentation accuracy around 90%.

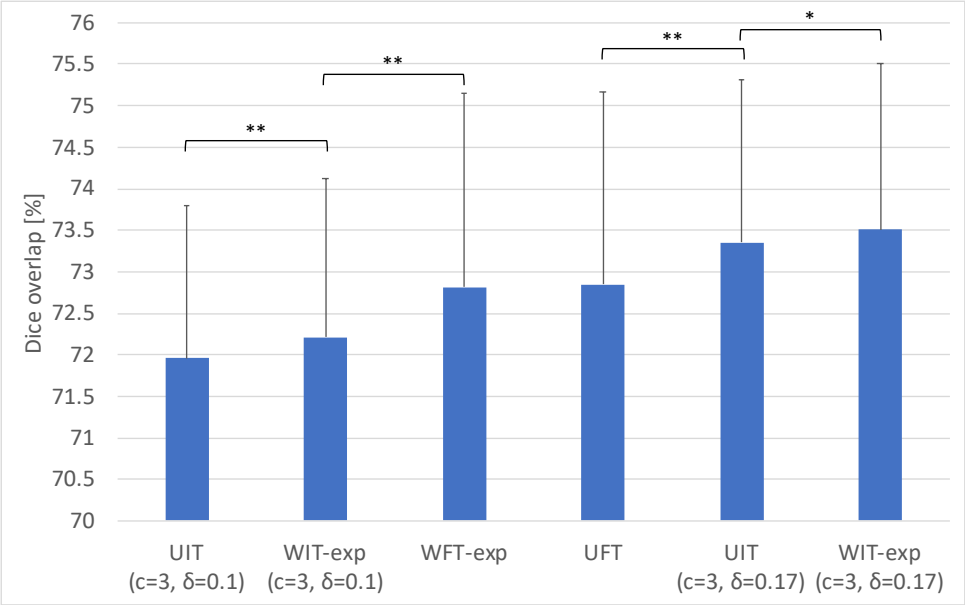


Figure 4.11: Comparison of fusion strategies and weighting schemes on all subsets of the MICCAI 2012 dataset with a TST. Error bars represent the standard deviation.

Out of 25 methods in the MICCAI 2012 ranking [125], this places us in the top 10 in terms of overall performance and in 5th position for non-cortical regions. Furthermore, our method required only 40 non-linear registrations at runtime for the 20 test images. This is in stark contrast to the other top 10 methods, including joint label fusion (74.99%), a variant of STAPLE (75.81%), MALP (75.76%) and STEPS (73.72%), which require at least 300 pairwise non-rigid registrations between atlases and targets at runtime.

4.10.6 Evaluation of fusion and thresholding strategies on the MICCAI 2013 dataset

The dataset was divided into 3 subsets of 12 images each and a TST was created for every target. In a cross-validation study we compared 4 fusion methods comprising unweighted fixed thresholding (UFT) and intensity thresholding (UIT), and weighted fixed thresholding and intensity thresholding with

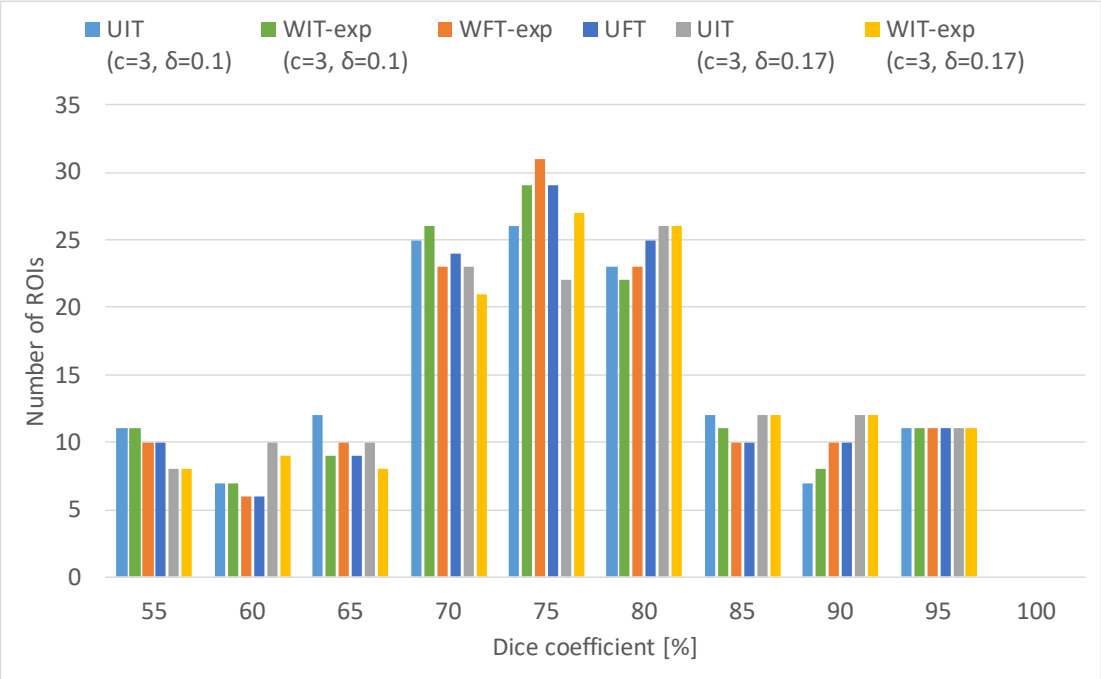


Figure 4.12: The histogram outlines the number of anatomical ROIs of the MICCAI 2012 dataset that fall within a certain range of accuracy.

exp weights (WFT-exp, WIT-exp). The fixed threshold θ was set to 0.35 for FT methods. For methods using IT, δ was set to 0.1 and the coefficient c to 3. We also tested WFT-exp with $\theta = 0.39$ since the noncortical regions of the dataset are characterised by a larger intensity distribution.

A significant improvement in segmentation accuracy was observed between UIT and WIT-exp ($p = 3.7 \times 10^{-5}$) and between UFT and WFT-exp ($\theta = 0.35$) ($p = 3.9 \times 10^{-4}$) which indicates the positive impact of using weights for both intensity and fixed thresholding (Fig. 4.13). Fixed thresholding with and without weights showed superior results compared to intensity thresholding, which might be related to the low intensity contrast in sub-cortical ROIs of the dataset. Raising the threshold for WFT-exp to 0.39 further improved accuracy significantly ($p = 1.1 \times 10^{-7}$). This is further supported by the comparison of fusion and thresholding strategies at regional level. Figure 4.14 shows that the main improvements of WFT-exp with different thresholds and UFT over UIT and WIT-exp were achieved

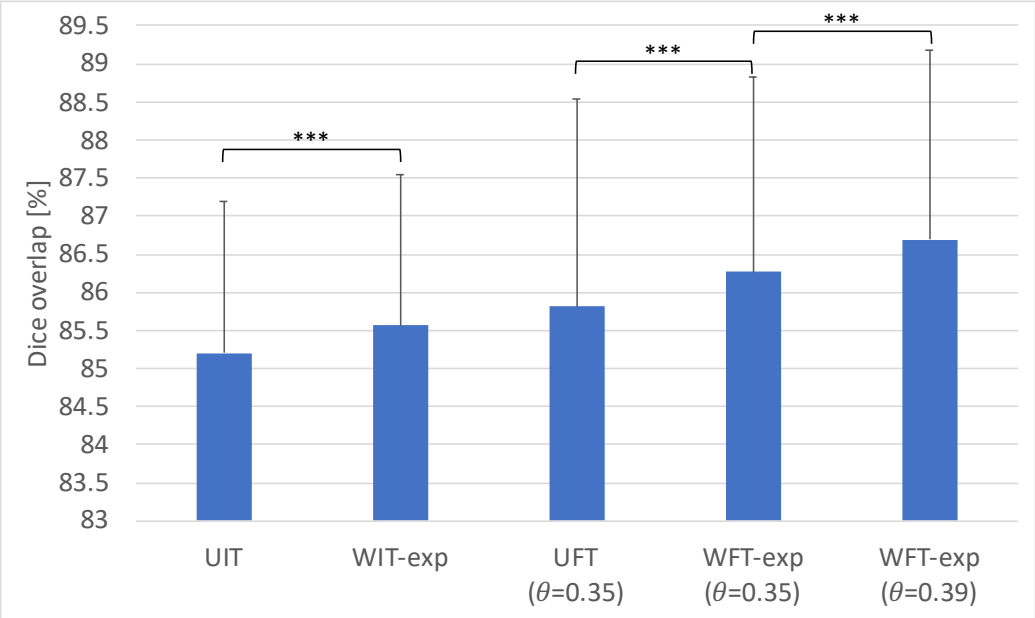


Figure 4.13: Comparison of fusion strategies and weighting schemes on all subsets of the MICCAI 2013 dataset with a TST. Error bars represent the standard deviation.

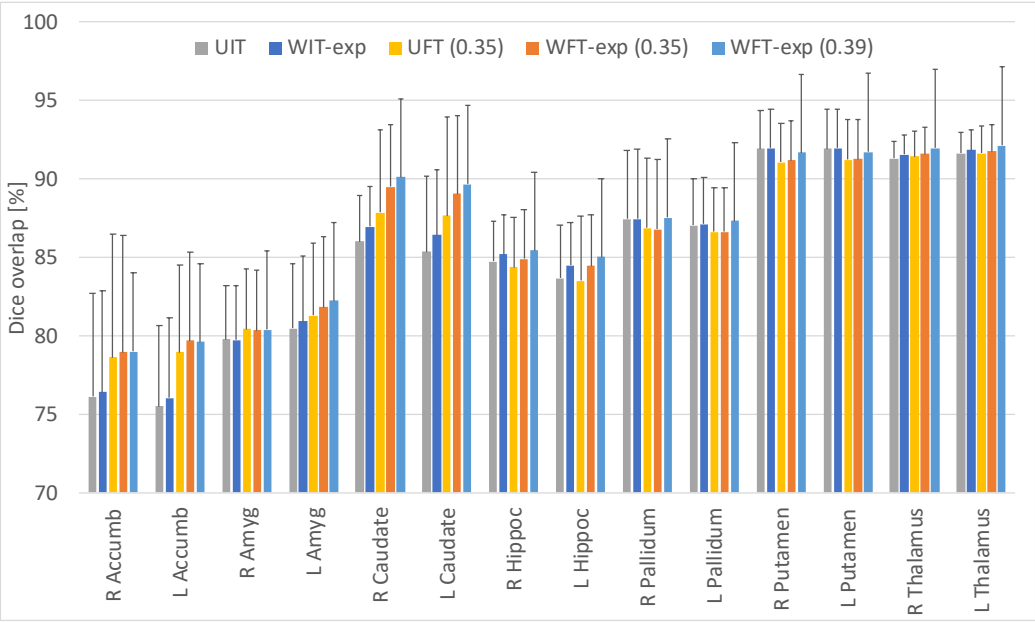


Figure 4.14: The histogram outlines the number of anatomical ROIs of the MICCAI 2013 dataset that fall within a certain range of accuracy.

in the Accumbens, Amygdala and Caudate ROIs, all of which have poorly defined borders.

Our best overlap result of 86.7 ± 2.48% was achieved with WFT-exp ($\theta = 0.39$), which is superior to Zikic *et al.'s* method [245], which achieved 82.47 ± 4.44% and Wu *et al.'s* method [236], which achieved 86.54 ± 2.59%. In [30] the authors implemented MV, STAPLE, STEPS and joint label fusion, which reached with the computationally most expensive SyN registration method 84.9 ± 4.2%, 84.4 ± 4.5%, 84.7 ± 4.9% and 87.5 ± 3.0% respectively. All except joint label fusion are inferior to our results. Their own method achieved slightly superior results of 87.1 ± 5.4% as reported in [31] and 86.5 ± 3.7%, 87.4 ± 2.9% and 88.0 ± 2.6% based on the registration method as reported in [30]. However, as mentioned previously in Section 4.10.2 the average population template would have to be constructed for each target at runtime, leading to enormous computational costs. The only alternative would be to directly register each atlas and target at runtime.

4.11 Discussion

The parameter settings for intensity thresholding were empirically determined on the LONI dataset by testing different values for the intensity distribution coefficient c and the threshold δ. Segmentation accuracy could be significantly improved by also considering the label probability. The parameters for intensity thresholding were adjusted to the requirements of the dataset for all following experiments. The fusion and thresholding was performed after construction of a TST for every target and propagation of the candidate labels to the target. In total 8 different combinations of fusion and thresholding strategies, i.e. MV, UFT, UIT, WFT-exp, WFT-1/sim, WIT-exp, WIT-1/sim, WLT and patch-refinement as an additional step, were tested on 6 datasets of different size and with different label maps.

Intensity thresholding approaches outperformed fixed thresholding on the LONI, ADNI-HarP and NIREP datasets, which pose very different chal-

lenges. For example, LONI's ROIs contain multiple tissue types and the hippocampus in ADNI-HarP has low intensity contrast. However, it requires the adjustment of the parameters to the labels of the dataset, which can be challenging when a label contains multiple tissue types and poor contrast to the surrounding regions, as found in the IBSR, MICCAI 2012 or MICCAI 2013 datasets. For example, fixed thresholding was superior than intensity thresholding with a small δ on the MICCAI 2012 dataset, which contains label maps with a large number of small cortical and non-cortical labels, making it difficult to set a constant δ. Increasing δ to sample intensity values from a larger region improved results for the NIREP dataset and a larger threshold θ showed better results for the MICCAI 2012 and MICCAI 2013 datasets. In general, the choice of δ depends on the intensity distribution in the ROIs. Smaller δ values showed good results for ROIs that contain the same tissue type and have a homogeneous distribution. Conversely, larger δ values achieved better results for ROIs that contain multiple tissue types and show a wider intensity distribution. Doshi *et al.* [74] additionally incorporated a tissue type categorisation term which compares the expected tissue type with the observed tissue type at a location, to make it more robust. An advantage of intensity thresholding is its capability to correct for misregistration, to some extent. If the intensity distribution of the target ROI is correctly sampled, small registration errors can be alleviated.

Our hierarchical approach with patch-based refinement was tested on the NIREP dataset. The patch-refinement was only applied to low-probability regions to provide a second set of weights, which allowed the re-estimation of the probabilistic label values on an even smaller scale. Considerable absolute changes in probability could be observed around high-probability regions, which provided a crisper and more accurate outline in the final segmentation, especially in the border regions. This was also evident by the statistical comparison of the Dice overlap measurements, which showed a significant improvement in segmentation accuracy. While high-probability regions were already segmented with high precision by our MAS approach, the applica-

tion of additional patch-refinement was only required in low-probability regions. Consequently, we could drastically reduce the computational burden of patch-based comparisons at runtime by approximately a half compared to pure patch-based algorithms.

Weighted fusion strategies outperformed unweighted strategies on the LONI, IBSR, MICCAI 2012 and MICCAI 2013 datasets. However, unweighted strategies reached similar results and slightly outperformed weighted strategies on the ADNI-HarP dataset. This could be caused by two reasons. First, we applied our similarity ranking and selection strategy for both types of approaches. Consequently, even the unweighted method used the highest ranked candidates for fusion. If all of the highest ranked candidates are very similar to the target, they are assigned a similar weight. Consequently, the weighting has less impact, which could result in a similar segmentation outcome to the outcome without using weights. In turn, if less atlases are provided the probability of finding candidates with a high similarity to the target might be smaller, leading to more extreme weights. This is also related to the weighting function. As seen in the LONI dataset 1/sim provided superior while exp provided inferior results. One alternative for the future would be to limit the number of candidates dynamically for each ROI. One approach has been developed by Rikxoort *et al.* [217], where the algorithm stops adding candidates when no more improvement can be expected. Second, our similarity measurement is based on the similarity of deformation fields. Low-intensity contrast regions might experience less deformation due to the lack of landmarks or intensity differences, which are used in the registration process. The similarity comparison would derive a ranking and weights from only small differences in distance in manifold space, which might not be reliable. Similarly, imprecise deformation fields could reduce accuracy of the weights.

Compared to other state-of-the-art methods, our approach has shown to provide competitive results, which places it amongst the top ranked methods for all of the six used datasets at a considerable reduction in computational

costs. Considering that we used established fusion methods for testing, there is potential for further improvements. We would expect higher segmentation accuracy when using our TST approach in combination with more sophisticated methods.

In general, the comparison of MAS methods is challenging due to the use of different registration methods, datasets and fusion strategies for evaluation. For example, the recently proposed method by Benkarim *et al.* [31, 30] has shown competitive results but was only tested on subcortical ROIs. In contrast, our method was applied to multiple diverse datasets with different cortical and non-cortical ROIs of different size. The MICCAI 2012 and 2013 MAS challenges provided a good platform for consistent evaluation. However they did not take computational efficiency into account. One of the most challenging datasets might still be the LONI-LPBA40, since it contains large ROIs consisting of white matter and grey matter without a clear anatomical landmark that would allow the distinction between parts of WM associated or not associated with a ROI. In the next chapter we will introduce a method to segment large anatomical regions of the brain into meaningful smaller ROIs, which allow a more localised candidate ranking and selection.

Chapter 5

Dynamically adjusted labels

5.1 Offline learning

Some tasks in the MAS framework, such as atlas selection, label propagation and fusion, can only be performed at runtime once the target image is given. However, information can already be gathered by studying the set of atlas images before the target is available, which is referred to as "offline learning". The collected information can then be used to enhance analysis of the target without affecting runtime. In the literature three main approaches to facilitate offline learning can be identified.

The first set of algorithms learns constraints, probability maps or features for classification from the atlas set, which can then be applied at runtime when segmenting the target. In the algorithm by Li *et al.* [134], a constraint on the ROI for image registration was enforced by detecting the anatomy of interest with regression forest classifiers to estimate a bounding box enclosing it. Van der Lijn *et al.* [216] proposed a method to build offline an intensity model from the atlas images, which provides the likelihood of a voxel intensity to be part of a particular label. In a related approach by Zikic *et al.* [245], one randomised classification forest for each atlas image is trained offline and allows the prediction of the label affiliation in the target without registering the atlases to the target. Instead, the spatial information about

label location is incorporated with a registered probabilistic population atlas, which reduces the computational costs. Methods that use appearance for voxel classification crucially depend on the similarity of the atlas and target intensity profile, but are generally faster than traditional MAS methods. Methods based on appearance have also shown promising results for the segmentation of highly varying anatomical shapes, such as tumors, where the mapping of corresponding structures is difficult [224].

Although image similarity between atlases and target is often used for atlas selection and fusion weight estimation, it is not necessarily related to the final segmentation accuracy. The second type of methods uses the set of atlases to predict their confidence in segmenting an image accurately or their weights for label fusion. Sdika [183] and Wan *et al.* [219] constructed accuracy or reliability maps to indicate the likelihood for every voxel to be correctly labelled by the atlas, or, in other words, how well the corresponding regions can be mapped. These maps can be obtained from the atlas images where the correct labels are known and later used as weights for the label fusion in a new target image. In Sjoeberg and Ahnesjoe's fusion method [189], the segmentation quality was assumed to be linearly related to the registration quality. The parameters for this linear relationship were estimated in a learning phase and later used to calculate the probabilistic weights for the fusion on a target image. Similarly, Sanroma *et al.* [180] learned the relation between image appearance and labelling performance to select atlases based on their expected performance rather than solely on image similarity.

The goal of the third type of algorithms is to allow fast selection of those atlases most similar to the target at runtime, for more accurate label propagation with potential TSTs as intermediate images. Instead of clustering atlases based on their appearance, Langerak [127] created clusters based on the quality of their deformation fields estimated for each pair offline. Other methods have used graph structures, where each atlas is represented by a

node, to find the most similar atlases to a target based on their distances to each other [110]. More sophisticated methods construct the graph in manifold space instead of image space, which has the advantage of reducing the high dimensional image space and providing more accurate distances between them [43, 44, 45, 75, 88, 234] (see Chapter 3).

Another task, which can be performed offline is ROI selection. While most methods use the pre-defined labels for local atlas selection and fusion, some ideas have evolved to either disregard the pre-defined labels and determine ROIs for TST construction based on the given atlases, or further parcellate the pre-defined ROIs of the atlases.

5.2 ROI selection

In Shi *et al.'s* method [187] ROIs were determined by parcellating the population template with a watershed algorithm (see Section 1.2.1.1) and clustering the set of atlases with affinity propagation [83] into sub-sets for each ROI. A probabilistic atlas was constructed for each of the sub-populations. When a new target is given, a TST can be rapidly assembled from those probabilistic atlases most similar to the target.

Rikxoort *et al.* presented an adaptive local MAS method [217] that requires two registration methods: A fast, computationally cheap linear method to align all atlases to the target and compute difference images, and a more accurate, computationally expensive registration that is only applied to those atlases with a low mean absolute difference. For atlas selection and label fusion, they further parcellated the given ROI into a pre-defined number of slightly overlapping and equally sized blocks. The number of accurate registrations could be reduced by introducing a criterion to automatically stop further registrations when the disagreement between the propagated labels falls below a certain threshold. Results showed that the local MAS variant could produce similar results to a conventional MAS method at reduced computational cost and when enforcing the same computational cost,

better results could be obtained.

An extension of the global STAPLE approach was presented in form of Spatial STAPLE [18], which accounts for spatially varying performance in form of a performance level field for each rater.

The advantages of local fusion strategies were also shown by Langerak *et al.*. In their local version of the SIMPLE algorithm [126, 128] ROIs were split into smaller sub-regions. They tested three different approaches: the division into a pre-defined number of blocks through the center of gravity, the division into user-defined regions, manually defined by an expert, and the division on a slice by slice basis, where each slice represents a sub-region. Their approach requires the registration of all atlases to the target and has an automatic stopping criterion for the fusion of each region based on convergence. Only the second approach which requires manual intervention for each dataset has shown significant improvement, which indicates that the selection of meaningful smaller regions has a positive effect on the weights and fusion. In general, these local selection and fusion strategies have shown to outperform their global counterparts [129, 229].
Although the automated division of segmentations into further local regions has shown improved results, this has been done without meaningful contextual (anatomical) information and manual division of pre-defined labels is influenced by inter- and intra-operator variability.

5.3 Clustering methods

While it has been shown that more localised ROIs can improve MAS, there is no automated method to divide given atlas labels into meaningful sub-ROIs. In our method we use clustering to determine sub-ROIs that undergo similar or dissimilar deformation. We will provide a review of clustering methods before outlining our approach.

5.3.1 K-means

One of the most popular unsupervised learning algorithms to classify data into a pre-defined number of clusters is k-means [82, 135, 137]. Given a set of n feature vectors, the goal is to classify the data into a pre-defined number of k clusters $C = \{C_1, C_2, ..., C_k\}$ with their respective mean values μ_i whereby $k \leq n$. The objective function for the j^{th} element x_j can be written as:

$$\underset{C,\mu}{argmin} \sum_{i=1}^{k} \sum_{x \epsilon C_i} \|x - \mu_i\|^2 \tag{5.1}$$

Lloyd's algorithm is usually initialised by randomly assigning a cluster to each element. Then it iterates through two main steps.. First, each element in the dataset is assigned to the cluster with the closest cluster centre (=mean of the cluster elements), hence, achieves the least sum-of-squares Euclidean distance. Second, after assigning each element to a cluster, the mean for each cluster is re-calculated to represent the new centre of the cluster. These steps are repeated until it converges and none of the elements change their assigned clusters. K-means clustering can also be described as a coordinate descent algorithm where the multivariate function is minimised along each direction at a time. One drawback of this approach is the need to randomly assign clusters to the elements, and consequently cluster centres, as an initial step. The outcome of the algorithm is sensitive to this first selection, and a good solution can only be obtained when the initial selection is close to it and a small number of clusters is used. In practice k-means is repeatedly applied with random initialisation values.

5.3.2 Affinity propagation

The k-means concept initially requires each element to be randomly assigned to a cluster, from which cluster centres are calculated. In contrast, affinity propagation [83] considers each element as a node in a network and initially represents a potential cluster centre. Instead of specifying a pre-defined number of clusters, the method takes the pairwise similarity measures between

the elements as input. A similarity measure between two elements x_i and x_k is calculated as $s(i,k) = -\|x_i - x_k\|^2$ and indicates the preference or affinity of x_i to consider x_k as a cluster centre. The selection process of choosing an element as a cluster centre depends on these so-called input preferences and additionally on the messages passed between elements. Two types of messages are exchanged. The first type of message sent from x_i to candidate centre x_k is called responsibility $r(i,k)$ and indicates how well x_k is suited as a centre for x_i after x_i considering all other potential candidates. The second type of message sent from x_k to x_i is called availability and indicates how suited x_k would be as a centre for x_i considering the responsibility sent from other elements to x_k. The method iterates through three main steps. First, the responsibilities are updated, keeping the availabilities fixed (Equ. 5.2). Second, availabilities are updated, keeping the responsibilities fixed (Equ. 5.3). Third, availabilities and responsibilities are combined to decide which elements should be the centres (Equ. 5.4). If the selection does not change for a certain number of iterations, the final solution is reached.

$$r(i,k) \leftarrow s(i,k) - \max_{k' s.t. k' \neq k} \{a(i,k') + s(i,k')\} \tag{5.2}$$

$$a(i,k) \leftarrow min \left\{ 0, r(k,k) + \sum_{i' s.t. i' \notin \{i,k\}} max\{0, r(',k)\} \right\} \tag{5.3}$$

$$a(k,k) \leftarrow \sum_{i' s.t. i' \neq k} max\{0, r(i',k)\} \tag{5.4}$$

5.3.3 Hierarchical agglomerative clustering

While the aim of methods such as k-means clustering and affinity propagation is to find similar elements to one representative exemplar of the cluster, hierarchical agglomerative clustering aims at partitioning the data where not all elements of a partition have to be similar to one central exemplar. Similarly to affinity propagation, each element is assigned to its own cluster and as the

algorithm proceeds pairs of clusters are merged. The merging procedure requires a metric and a linkage criterion to decide which clusters are combined. A similarity metric such as the Euclidean distance (Equ. 5.5) can be used to quantify the pairwise distance between elements a_i and b_i. When the data contain mixed attributes with large differences in their variance, the Mahalanobis distance (Equ. 5.6) or standardised Euclidean distance (Equ. 5.7) can be used to scale the differences between the coordinates of the elements.

$$\|a - b\|_2 = \sqrt{\sum_i (a_i - b_i)^2} \tag{5.5}$$

$$\sqrt{(a-b)^T S^{-1} (a-b)}, \quad \text{where S is the Covariance matrix} \tag{5.6}$$

$$\sqrt{\sum_{i=1}^{N} \frac{(a_i - b_i)^2}{s_i^2}}, \quad \text{where s is the standard deviation} \tag{5.7}$$

The linkage criterion is a function of the pairwise distances within a cluster and quantifies the similarity between clusters. Commonly used methods for computing the distance between clusters A and B are based on the average (Equ. 5.8), centroid (Equ. 5.9) or shortest distance (Equ. 5.10).

$$\frac{1}{|A|\,|B|} \sum_{a \epsilon A} \sum_{b \epsilon B} d(a, b) \tag{5.8}$$

$$\|c_s - c_t\| \tag{5.9}$$

where c_s and c_t are the centroids of clusters s and t, respectively.

$$min\,\{d(a, b) : a \epsilon A, b \epsilon B\} \tag{5.10}$$

5.4 Our approach to label adjustment

5.4.1 Dividing the labels into clusters

In order to divide larger labels into meaningful smaller sub-ROIs, we warp all atlas labels $\{\widetilde{L_j^r}\}_j$ on the population template with their associated deformation fields $D_{\widetilde{I}_j \to \overline{I}}$ resulting in a set of overlapping segmentations on the population template $\{\overline{L_j^r}\}_{j=1...N}$. If the estimated deformation fields were accurate, these would match perfectly. As this is usually not the case, we calculate for each ROI the union of the corresponding warped atlas labels, which outlines the largest possible area covered, i.e. $C_r = \{x | \exists j \in [1, N], \overline{L_j^r}(x) > 0\}$. At the same time we calculate the standard deviation of the deformation magnitude of each voxel of the atlas deformation fields resulting in a map S that shows areas of varying degree of deformation (Fig. 5.1). In order to dynamically divide the pre-defined areas, given as C_r, into sub-ROIs we apply an agglomerative clustering algorithm to each of the extracted areas $V_{\widetilde{I}}^r = \{S(x) | x \in C_r\}$ of this map and their voxel position. Voxels close to each other with a similar standard deviation will be in the same cluster and different clusters indicate regions of varying degree of morphological preservation in the image dataset. Within the same token we create a table, which provides the correspondence between pre-defined labels and clusters, which is later used for the fusion. Each of the clusters is projected back on the atlases with the inverse of the respective deformation field. As the unions C_r are generally larger than the individual labels, the back projected clusters are cropped with the pre-defined labels. From this point on the clusters can be used in the same way as the original labels for TST construction and propagation to the target.

Next to smaller sub-ROIs for more localised candidate selection and fusion, our approach also holds the potential to estimate the ideal number of candidates for each cluster. Well preserved regions with a low variability in the dataset could consult more candidates for fusion while regions with a high variability could consult a smaller number of candidates to introduce

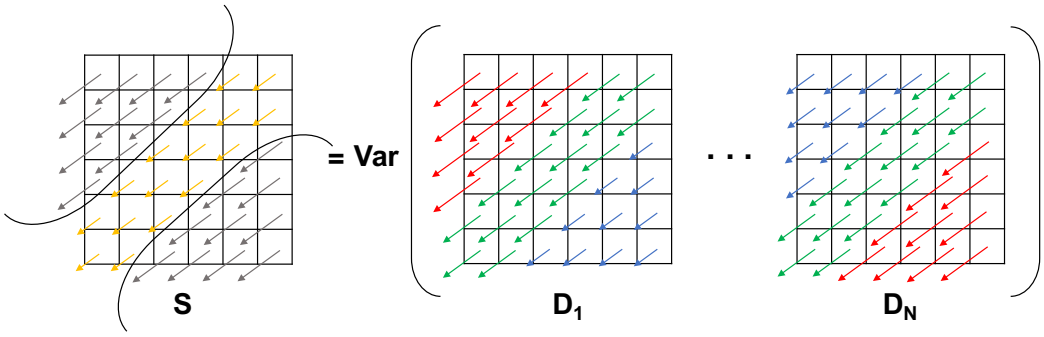


Figure 5.1: Schematic representation of our approach to divide ROIs into meaningful sub-ROIs. We calculate the variance of the deformation for each voxel from the individual deformation fields, in which red indicates large, green indicates medium and blue indicates small deformation. Hierarchical clustering is performed on the magnitude of the deformation vector and the locations of the voxels. The map S of resulting clusters characterises three sub-ROIs of different degrees of variance (grey=large variance, orange=small variance).

less errors from dissimilar atlases.

5.4.2 Combining the clusters

Once the clusters have been propagated to the target they have to be re-combined into the pre-defined anatomical labels. We use the correspondence table to find the corresponding label each cluster belongs to. We start an iterative process by fusing the first, which is the highest ranked, candidate of each cluster. To make sure that this first layer is continuous and smooth, the mean is calculated in overlapping regions between clusters. This is repeated for the next highest ranked candidate of each cluster, which creates another layer. Once the maximum number of candidates is reached, we calculate the average for every voxel over all layers. Similar to the fusion methods in Section 4.9, this fusion can be performed weighted or unweighted.

5.5 Experiments

5.5.1 Evaluation of the clustering method on the ADNI-HarP dataset

The dataset was divided into 5 subsets of 26 images each. In a cross-validation experiment we compared the segmentation accuracy achieved with 3 different combinations, without using a TST or clustering (no TST, no cluster), with a TST but no clustering (TST, no cluster) and with both a TST and clustering (TST & cluster). WFT-exp and WIT-exp were used for the fusion and thresholding of 5, 10, 15 and 40 candidates. The fixed threshold θ was set to 0.35 and the intensity thresholding parameters to $\delta = 0.1$ and $c = 3$. One subset of 7 target images was used to assess the impact of varying cluster numbers. Each of the two ROIs was divided into 4 and 8 clusters and the candidate fusion was performed with WIT-exp of varying candidate numbers. Note that for this comparison we also calculated and applied weights to the variant without a TST and clusters.

The division of each ROI into 8 clusters showed improved segmentation accuracy over 4 clusters (Fig. 5.2) with all candidate choices of 10, 15 and 40 candidates. Note that due to the similar label sizes of the left and right hippocampus they could be divided into the same number of clusters. The labels of other datasets can vary considerably in size and require special care when the number of clusters is chosen.

With WFT-exp the use of our TST and our clustering method improved segmentation accuracy for all candidate choices (Fig. 5.3a). With regards to the candidate choices, the biggest improvement was achieved when using 10 instead of 5 candidates, which reduced when more candidates were added. The use of a TST and clusters with 5 candidates reached a Dice overlap of 82.99 $\pm$ 4.03% and outperforms the 82.8 $\pm$ 4.26% with 40 candidates, no TST and no clusters. Consequently, the use of a TST and clusters is more efficient than traditional MAS methods as it requires less candidates to reach

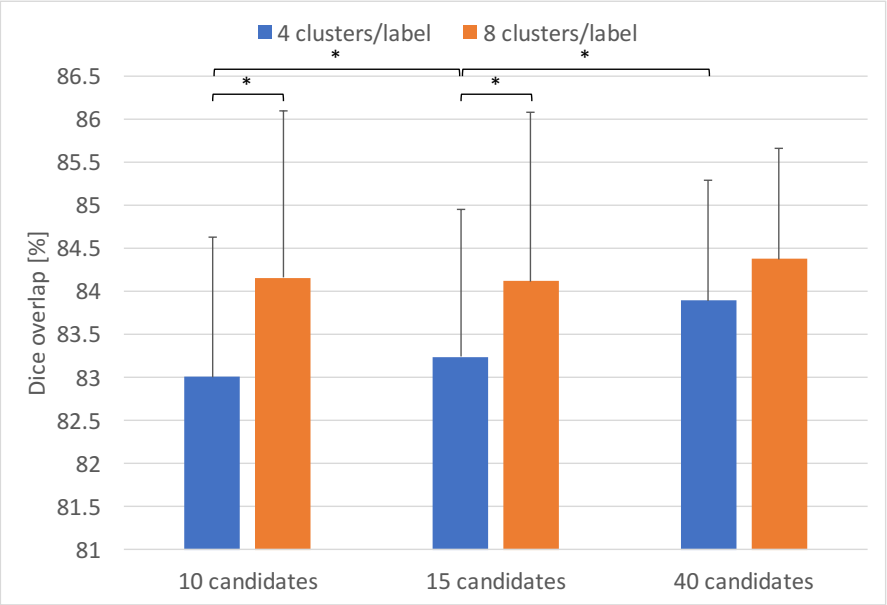
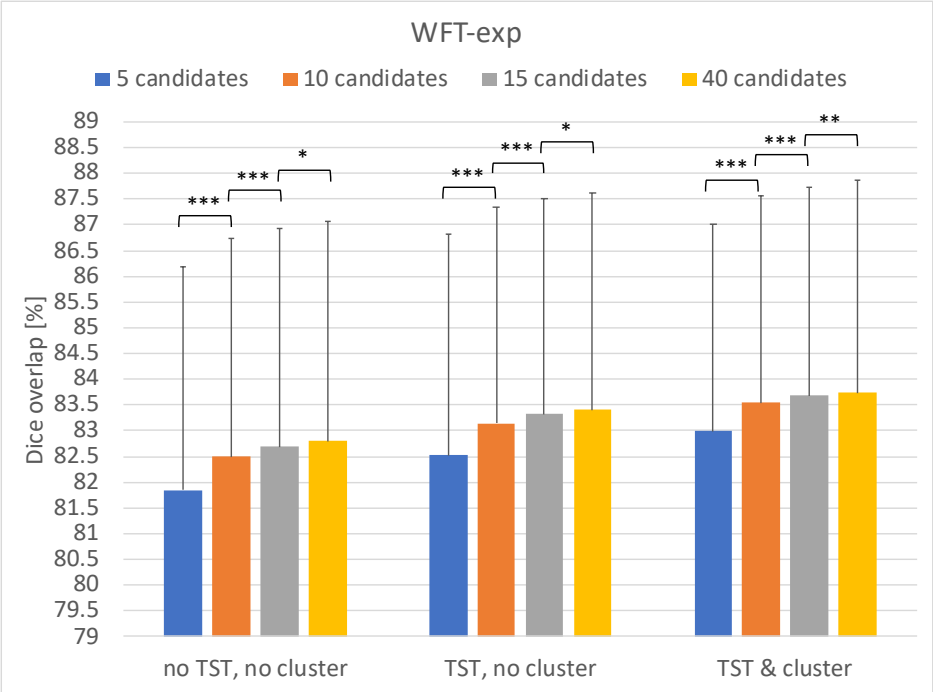


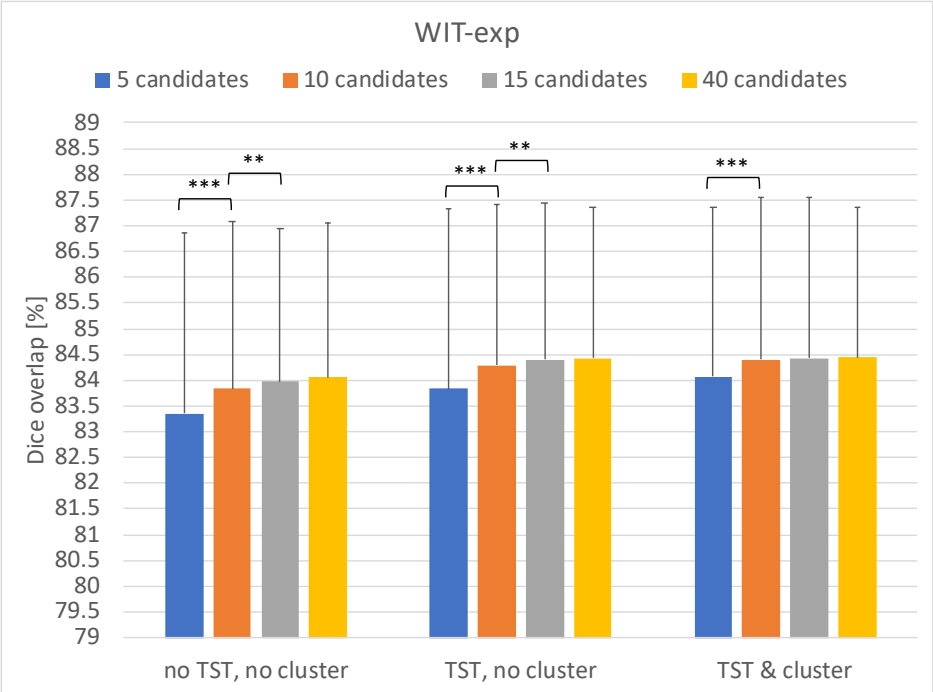
Figure 5.2: Effect of changing the number of clusters from 4 to 8 clusters per label with WIT-exp of 10, 15 and 40 candidates on a subset of 7 targets of the ADNI dataset. Error bars represent the standard deviation.

or even exceed the same level of accuracy.

The use of WIT-exp showed better results than WFT-exp for all candidate choices and variants (Fig. 5.3b). The use of clusters showed improvement with a small number of candidates. With increasing number of candidates the use of clusters had less impact. With regards to the number of candidates, the biggest gain was observed when using 10 instead of 5 candidates and only little to none with more candidates. This indicates that the intensity thresholding method can mitigate segmentation errors but does not improve results beyond a certain level. By using a TST and clusters, it required only 5 candidates to obtain a level of accuracy similar to that achieved with 40 candidates, when no TST and no clusters were used.



(a) WFT-exp



(b) WIT-exp

Figure 5.3: Improvement in segmentation accuracy by using a TST and clustering with a) WFT-exp and b) WIT-exp of 5, 10, 15 and 40 candidates on all images of the ADNI dataset. Error bars represent the standard deviation.

5.5.2 Evaluation of the clustering method on the MICCAI 2013 dataset

The dataset was divided as outlined in Section 4.10.6, a TST was constructed for every target and the same 4 fusion methods and parameters were applied. Similarly to the previous experiment, the smallest ROI was divided by a fixed factor. The remaining ROIs were divided into multiples of the factor depending on their size. First, we evaluated the use of different numbers of clusters on segmentation accuracy on all test sets. Second, we evaluated the impact of different numbers of candidates and fusion methods on the segmentation accuracy of one testing set. Third, we compared the overall performance to the use of a TST alone and no TST.

The impact of a different number of clusters on segmentation accuracy is illustrated in Figure 5.4. We divided the initial 14 ROIs into a total of 53 and 106 clusters. A significant improvement was achieved when using 53 clusters for intensity-thresholded methods ($p<.001$) while, the use of 106 clusters showed significantly higher accuracy for WFT-exp ($p=0.043$).

The use of only 15 candidates for fusion compared to all candidates improved segmentation accuracy significantly with UIT ($p=7.1 \times 10^{-3}$) (Fig. 5.5). No significant difference was found with WIT-exp, UFT and WFT-exp.

Segmentation accuracy significantly decreased for UIT and WIT-exp with the use of clusters ($p<.001$). However, a significant positive effect was observed in combination with fixed thresholding methods ($p<.01$) (Fig. 5.6).

5.5.3 Evaluation of the clustering method on the NIREP-NA0 dataset

The dataset was divided into 6 subsets of 3 images each. In a cross-validation experiment we compared the segmentation accuracy achieved with 3 different combinations, without using a TST or clustering (no TST, no cluster), with a

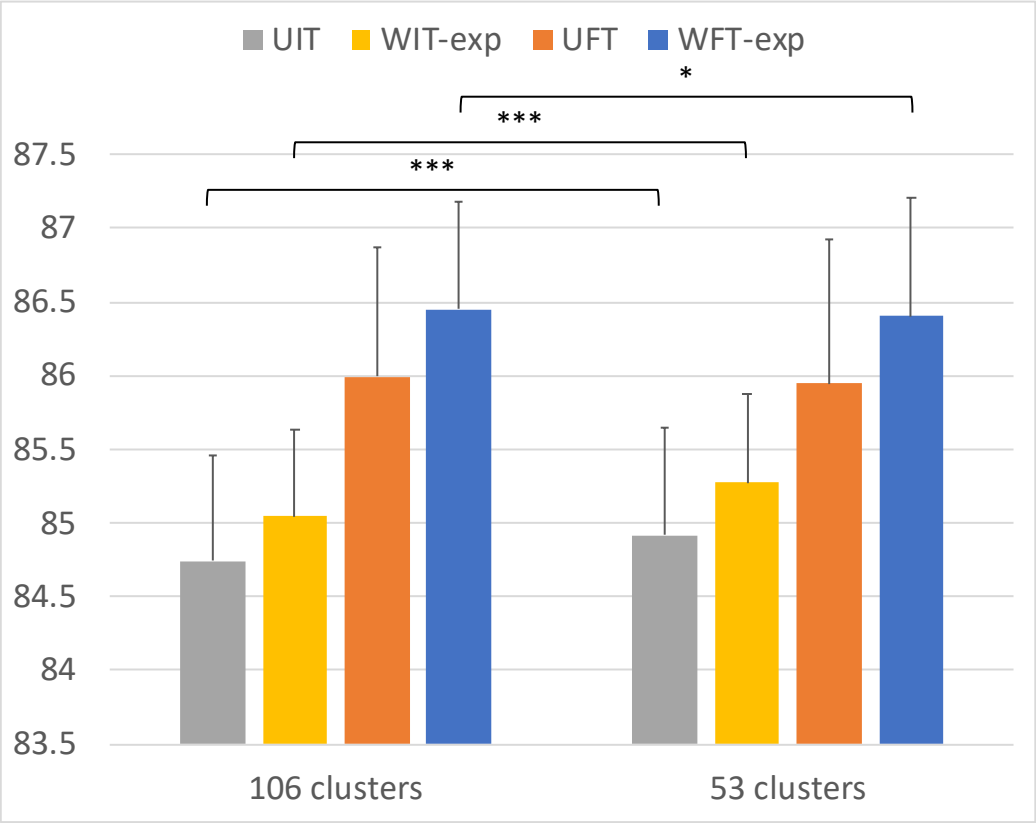


Figure 5.4: Evaluation of different numbers of clusters on segmentation accuracy on the MICCAI 2013 dataset. Error bars represent the standard deviation.

TST but no clustering (TST, no cluster) and with both a TST and clustering (TST & cluster). Due to the varying label size, the number of clusters was dynamically determined for each label. We used the smallest label size as a reference and divided it into 2 clusters. The sizes of the remaining labels were divided by the the size of the smallest label and the number of required clusters calculated as a multiple of 2. The fixed threshold θ was set to 0.35 for FT methods. For methods using IT, δ was set to 0.1 and the coefficient c to 3. No patch-refinement was used.

Accuracy significantly increased with the clustering method and WFT-exp (Fig. 5.7). However, no significant effect could be found for WIT-exp.

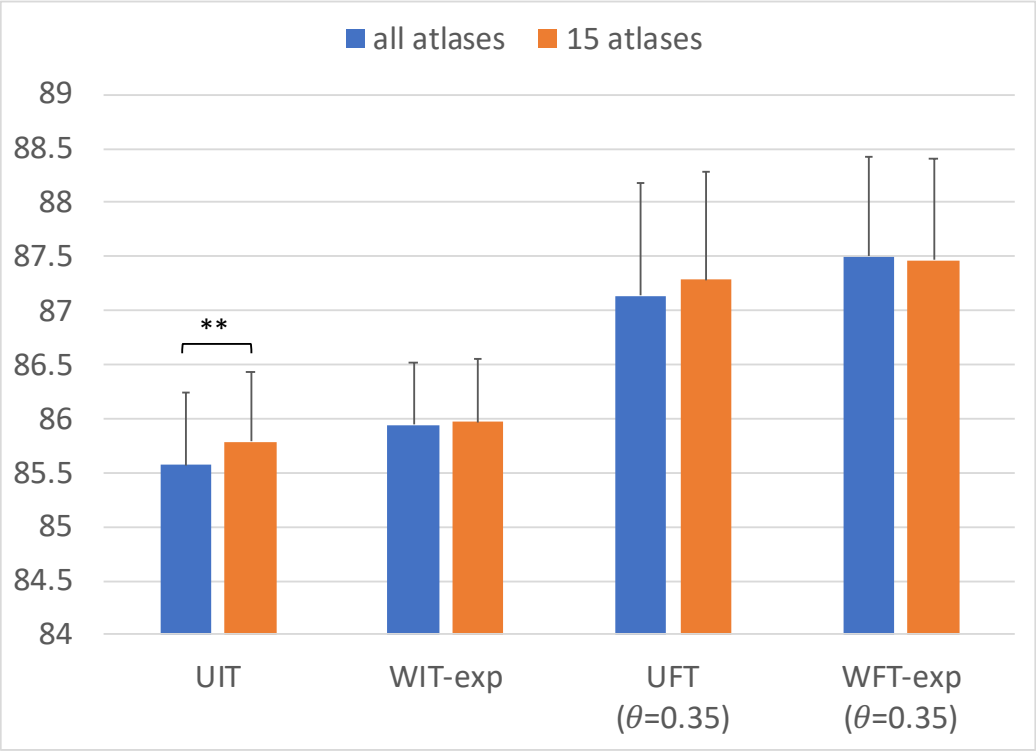


Figure 5.5: Evaluation of different numbers of candidates on segmentation accuracy on one subset of the MICCAI 2013 dataset. Error bars represent the standard deviation.

5.6 Discussion

The use of our clustering method in addition to the TST further improved segmentation accuracy of the ADNI dataset with both fusion strategies, and the NIREP and MICCAI 2013 dataset with the fixed thresholding methods. However, clustering showed a negative effect with the intensity thresholding method on the NIREP dataset, which might be because intensities are already sampled from a representative region of the label without clusters. The clustering showed a negative impact for intensity-based thresholding methods on the MICCAI 2013 dataset. Although the use of clusters improved accuracy with fixed thresholding and the chosen threshold of 0.35 on the MICCAI 2013 dataset, we achieved better results without clustering and a threshold of 0.39. This suggests that, similar to the intensity-based methods,

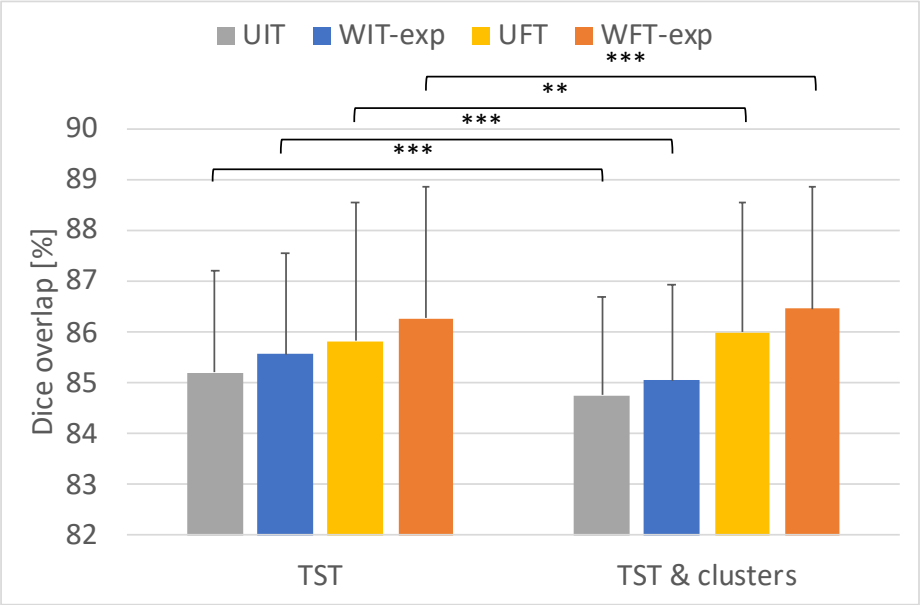


Figure 5.6: Evaluation of different fusion methods and the use of a TST on segmentation accuracy on the MICCAI 2013 dataset.

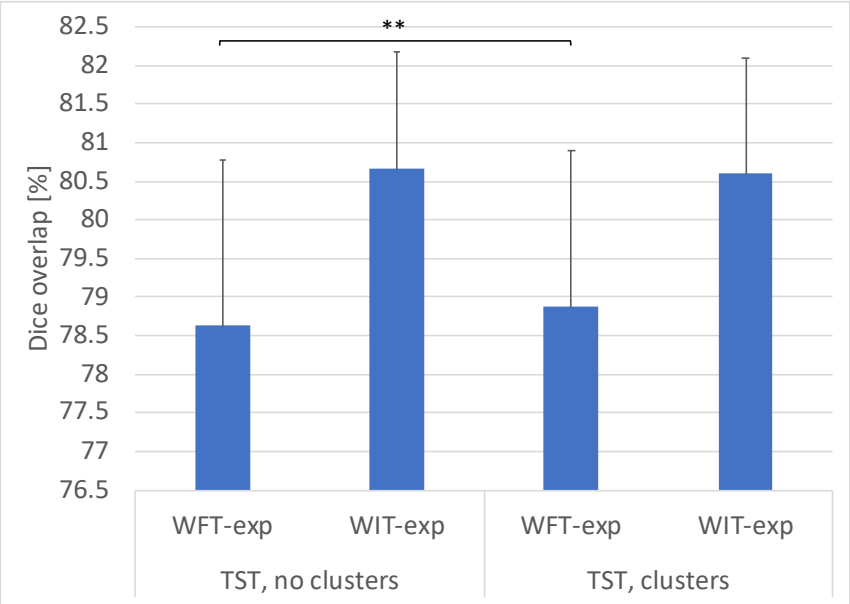


Figure 5.7: Improvement in segmentation accuracy by using a TST and clustering with WFT-exp and WIT-exp with all candidates and images on the NIREP dataset.

the parameter configuration requires adjusted to the labels of the dataset.

Increasing the number of clusters improved results of the ADNI dataset for all thresholding methods. For the MICCAI 2013 dataset increasing the number of clusters had a positive effect for fixed thresholding but a negative effect for intensity thresholding. Since the ROIs of the ADNI dataset outline the corresponding anatomical structures in the left and right hemisphere they are approximately of same size. However, it is more complicated with other datasets where label sizes vary considerably. For example, the IBSR dataset includes a large label for each cortical hemisphere, but also small labels like the ventral diencephalon. Although the clustering is performed offline and has to be done only once, a large number of clusters can have a negative impact on the runtime, because every candidate label has to be warped to the target. In our experiments with the NIREP dataset we chose to use the smallest label as a reference and divided it into 2 clusters. The number of clusters for the remaining labels was calculated as a proportional multiple thereof. If applied to the IBSR dataset, this could result in approximately 140 clusters just for the left hemisphere label. Consequently, a trade-off between accuracy improvement and runtime has to be found. Another strategy would be to use small labels without clustering and only divide larger labels. This might explain why other authors chose to evaluate their local MAS methods on large single-label anatomies such as the heart [217]. Due to the already large number of labels and their small size, the MICCAI 2012 dataset was not used for our evaluation. Similarly to results from the literature, our method can drastically reduce the number of necessary candidates and, consequently, computational cost, to achieve the same or even better performance compared to traditional MAS methods.

In the future we will test our approach with different clustering methods such as affinity clustering, which is not parametrised by the number of clusters, and more sophisticated fusion strategies. Although the similarity between the target and candidates is usually encoded in the fusion weights, sub-ROIs describing varying degrees of consistency in the data might allow

us to predict the ideal number of candidates for fusion. For sub-ROIs with a high variability, a small number of similar candidates might be beneficial to minimise the introduction of errors by unsuitable candidates. Conversely, for sub-ROIs with a small variability more candidates could be taken into consideration.

In conclusion, we think that considering the increasing amount of publicly available datasets of healthy and diseased cohorts, which could be potentially used as atlases, shifting the computational burden offline becomes more important. Although large datasets hold a great potential for MAS, they also require the delineation of anatomical structures to be able to employ them as atlases. One way to collect large amounts of labeling information is by using non-expert opinions. Another way is to improve algorithms to learn from corrupted data. Examples to gather labeling information for larger datasets include the approach by Bryan *et al.* [40], which allows users to assess their performance, and Bogovic *et al.* [36], which provides a detailed protocol for inexpert human raters to assign labels to anatomical structures of the cerebellum. While this concept might not be applicable to every anatomical ROI and may depend on the quality of the MRI scans, it is certainly an interesting direction for the future.

Chapter 6

Application to Tourette's images

6.1 Introduction

Although this Ph.D. thesis focuses on the development of an automated segmentation method for MR brain scans, it was part of a larger project to investigate differences in anatomical structures of adolescents with Tourette's syndrome (TS) compared to healthy subjects. With the James Tudor Foundation as the funding organisation and our presence in the school of psychology we were able to test our approach on an in-house acquired dataset of healthy and TS subjects.

6.1.1 Healthy brain development

Through its lifetime, from a basic prenatal stage on to maturation in adulthood and later decades of life, the human brain undergoes a series of structural alterations. Although the development is a continuous ongoing process and developmental stages can vary between subjects, some critical key events can be pointed out. In the very early embryo, approximately 3 weeks after conception, the neural tube develops out of reformed ectodermal tissue. All further parts of the nervous system then differentiate from this initial struc-

ture. From week 5 of gestation components like the forebrain, ventricles and spinal cord begin to evolve. The proliferative zones next to the ventricles differentiate and new neurons grow and migrate to their prospective components of the brain until around week 20. Simultaneously, the main sulci and gyri start to develop around week 15, with the major sulci already in place by 28 weeks of gestation. Between week 24 of gestation and week 4 after birth apoptosis sets in and about half of the cells die. Axon myelination starts around week 29 and while some areas are already myelinated by the onset of childhood, some parts of the brain finish the process in late adulthood. The myelinated axons build the white matter and the cell bodies of the synapses are mainly in the grey matter. Meanwhile, synaptogenesis starts around the 20^{th} gestational week and leads to a rapid increase in the formation of synapsis and, consequently, synaptic density reaching a maximum value in the frontal cortex 1-2 years after birth. After a plateau phase from 2 to 7 years, when synapse formation and cell deaths are counterbalanced, a decrease in the number of synapses can be seen. This complex developmental process leads to rapid brain growth from birth till age 2 years, by which it has already reached 80% of its full-grown adult weight [142].

The development is an ongoing process and the structure and density of grey and white matter changes continuously over a human's lifetime. The development of the cortical grey matter volume (considered as a whole) shows a parabolic (inverted U) pattern, but the different lobes of grey matter peak at different stages of development. For example, the frontal lobe grey matter peaks at 12.1 years in boys and 11.0 years in girls, while the temporal lobe cortical grey matter peaks at 16.2 years in boys and 16.7 years in girls. In contrast to the pattern seen in grey matter, the volume of white matter shows an increase throughout the development in childhood and adolescence [133].

Because of the extraordinary complexity of the brains growth process, it is very susceptible to disruptions, especially in fetal and postnatal stages. Disturbances due to individual genetic abnormalities and environmental factors in this critical period might have permanent consequences and can lead

to lifelong disabilities [93].

6.1.2 Tic disorders and Tourette's Syndrome

According to the ICD-10 version (WHO, 2010) symptoms (occurrences, events) such as involuntary, rapid, recurrent, non-rhythmic motor movements or sudden vocal productions that serve no purpose are considered as tics and classified as behavioural and emotional disorders with onset in childhood and adolescence. Tics include eye-blinking, neck-jerking, shoulder-shrugging, throat-clearing, barking, sniffing, and hissing and, although they can usually be supressed for a while, they inevitably appear at some point. Tic disorders can be divided into transient disorders that persist for less than 12 months and chronic disorders, either motor or vocal, for more than 12 months. In general, tic disorders are much more common in boys than girls.

Patients with TS show combined manifestations including both vocal and multiple motor tics that do not have to necessarily occur concurrently. Although severity and frequency of tics can vary substantially between distinct tic disorders and under different conditions they are commonly seen as parts of the same spectrum with transient disorders at the one end and Tourettes at the other end. Usually tics begin to appear during early childhood or adolescence (2-15 years) but intensity, distribution and characteristics can change over time. The peak is most commonly reached during early adolescence and from then on a decline or even disappearance can be seen [182]. While stressful situations can aggravate severity of tics, they disappear during sleep (WHO, 2010; APA, 2013). The prevalence was underestimated for a long time and is now thought to be around a rate of 0.3% to 0.8% of the school-age population. Patients with TS often have comorbidities such as ADHD and OCD [182].

In adults, cortical thinning was found in the sensorimotor and premotor cortex [233]. In pediatric TS subjects a thinning of the cortex was found in the sensorimotor, frontal, parietal and occipital cortex [197]. Functional neu-

roanatomical studies have highlighted brain regions whose activity is highly correlated with tic behaviour. The regions include the premotor cortices, anterior cingulate cortex, prefrontal cortex, parietal cortex, putamen, caudate nucleus and primary motor cortex [198].

Another research focus has been on subcortical regions, due to their importance in movement disorders and key roles in learning and motor control. Smaller caudate nucleus volumes were found in children and adults with TS [156] and an inverse correlation was found between the reduced volume of the caudate nucleus in childhood and the increased symptom severity in early adulthood, although no correlation was observed at the time of the MRI scans in childhood [34]. However, the comparison of multiple studies of subcortical region volumes has shown mixed results. Smaller left putamen and pallidus volumes were observed in adults and children [154, 156], while enlarged thalamus [132], amygdala and hippocampus [155] volumes were found in children. However, smaller hippocampus and amygdala volumes were found in adults [155]. It was suggested that this discrepancy could be caused by the use of different methods for analysis, which are generally less advanced for subcortical ROIs compared to cortical ROIs [52]. Consequently, to study the developmental process of the brain in patients with TS and to find differences with respect to the healthy developing brain, data throughout the disease progression and accurate medical image analysis methods for segmentation are required.

6.1.3 Method

We used the MICCAI 2013 dataset for training and the method described in Chapter 3 to segment 14 sub-cortical structures in scans of 27 subjects with TS and 27 age-matched healthy controls (HC), aged 8 to 21 years. WFT-exp was used for fusion with the same parameters as outlined in Section 4.10.6, where it showed the overall best performance compared to other tested fusion strategies.

The statistical analysis was performed on the regional, volumetric results of the segmentation corrected for intracranial volume. For each region we performed a cross-sectional comparison of all TS and HC subjects, and paired t-tests to find significant differences. A repeated measures general linear model with age as a covariate was used to compare the growth trajectories for each ROI.

6.2 Results

The cross-sectional, region-wise comparison for the two groups (Fig. 6.1) showed no significant difference between HCs and subjects with TS.

We found a statistically significant difference in the growth pattern of the right pallidum ($p<.05$). For all other ROIs no significant difference was found. A similar growth pattern in HCs and TS subjects was observed in the accumbens, caudate and right putamen (Fig. 6.2-6.3). In the amygdala, hippocampus, pallidum, thalamus and left putamen the volume in TS subjects was reduced compared to HCs in the younger population and increased with age. For HCs the volume increased at a slower rate in the amygdala and hippocampus compared to TS subjects. The volumes of the left pallidum, the right putamen and the left thalamus in HCs stayed constant and the right pallidum, left putamen and right thalamus decreased with age.

6.3 Discussion

In our cross-sectional analysis we did not find significant volumetric differences. More adolescent subjects would be required to look at age-specific differences. We could not replicate the findings reported by Peterson *et al.* [154, 156] where they found smaller left pallidum and putamen volumes in adults with TS. On the contrary, our results showed a tendency towards enlarged volumes in the right pallidum, left putamen, left amygdala, hippocampus and thalamus in adults with TS and smaller volumes in children.

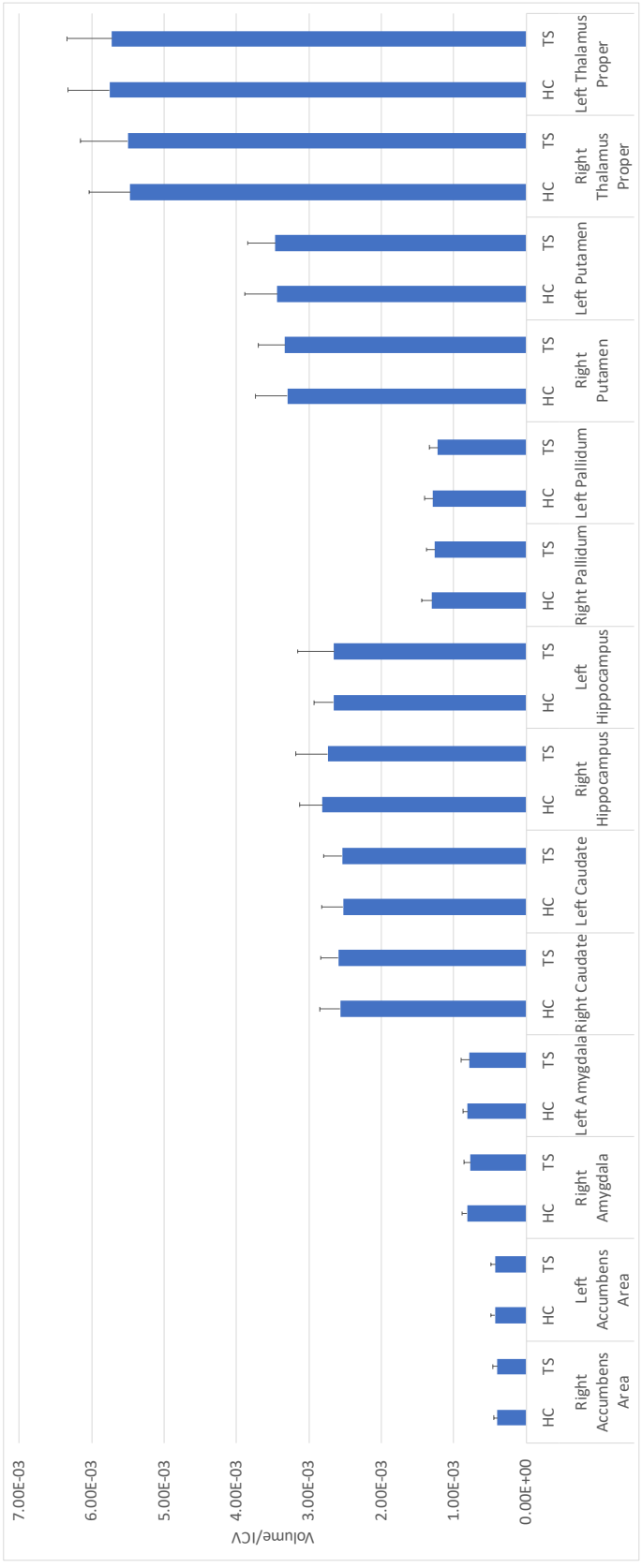
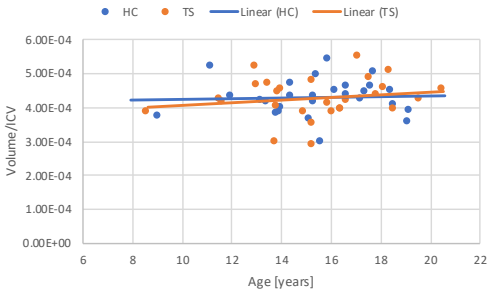
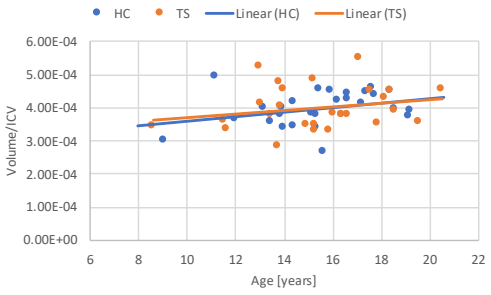


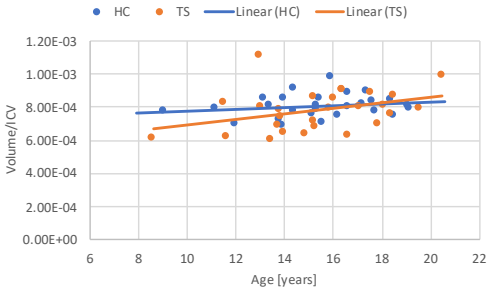
Figure 6.1: Volume comparisons of 29 HC and 29 TS subjects. Error bars represent the standard deviation.



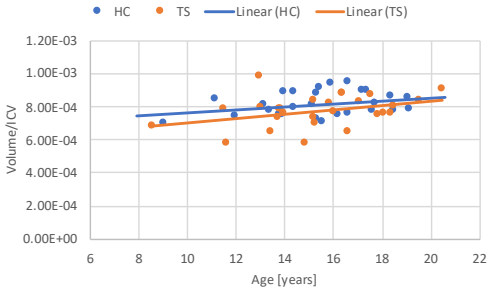
(a) Left Accumbens



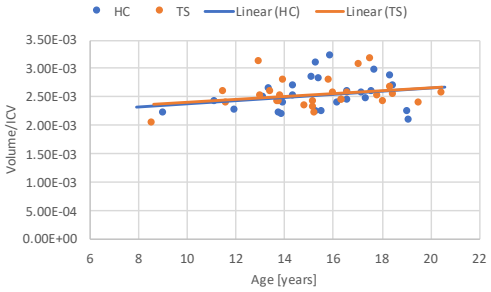
(b) Right Accumbens



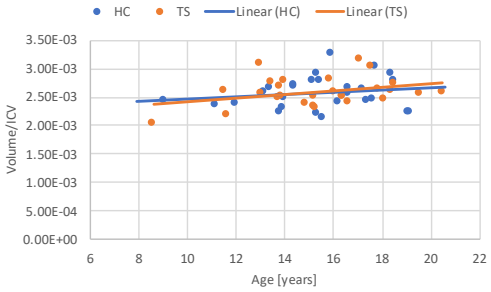
(c) Left Amygdala



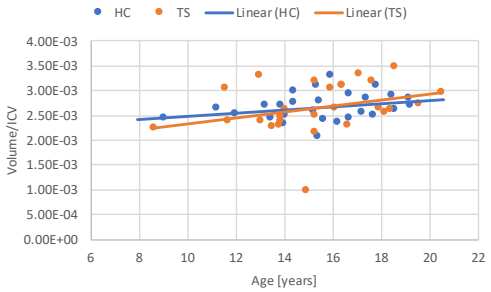
(d) Right Amygdala



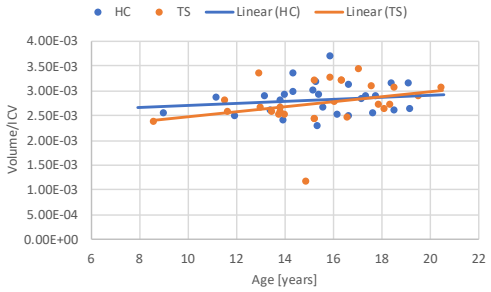
(e) Left Caudate



(f) Right Caudate

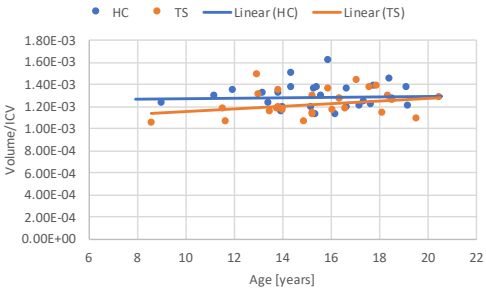


(g) Left Hippocampus

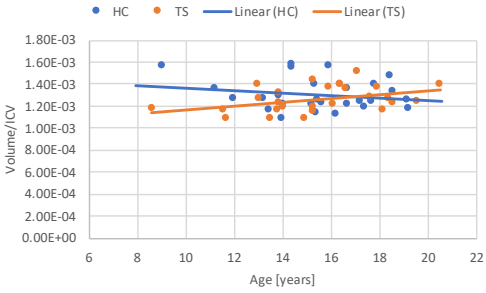


(h) Right Hippocampus

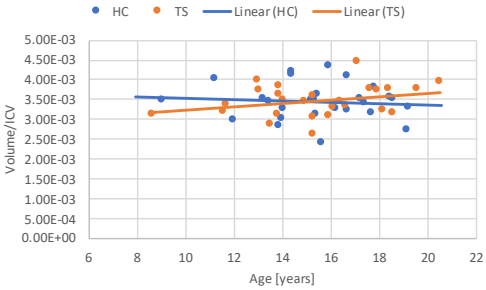
Figure 6.2: (1/2) Change in volume of sub-cortical ROIs in healthy controls (HC) and Tourette's subjects (TS).



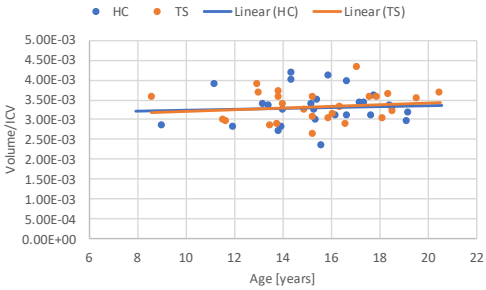
(a) Left Pallidum



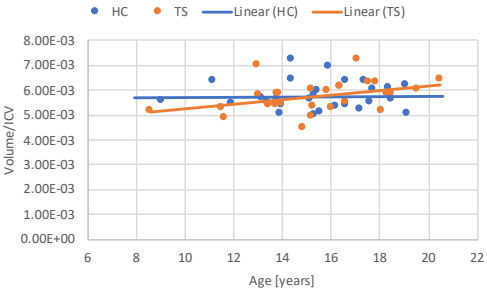
(b) Right Pallidum



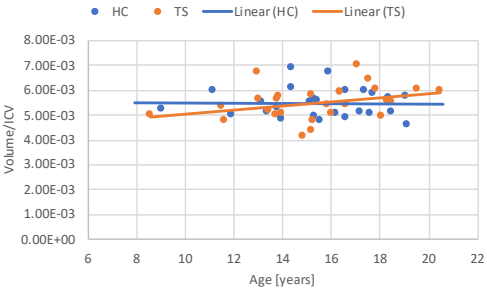
(c) Left Putamen



(d) Right Putamen



(e) Left Thalamus



(f) Right Thalamus

Figure 6.3: (2/2) Change in volume of sub-cortical ROIs in healthy controls (HC) and Tourette's subjects (TS).

This is in line with several other studies [132, 155, 165] where enlargement in the thalamus, amygdala, hippocampus and putamen in adults was found. In the developmental process of HCs the volumes of the thalamus, the right putamen and the left pallidum and accumbens were constant while a change was noticeable in TS subjects. The caudate volumes and change in volumes of HCs and TS subjects was similar, which is in stark contrast to Peterson *et al.'s* findings [156], where smaller caudate volumes were observed in children and adults with TS.

One drawback of our experiment was the use of the MICCAI 2013 dataset as atlases, which contains images from a wide range of ages (18-90 years, mean=32 years). A manually labelled public dataset for the slightly younger age range of our populations was not accessible but might be even more accurate with our MAS approach. The sample size might be too small to find an effect between the regional volumes of HC and TS subjects. Increasing the sample size would allow the analysis at different developmental stages in adolescents in more detail.

While most studies have focused on children or adults, our findings provided an estimate of the growth pattern and volumetric differences of subcortical structures in adolescents, which could complement current results.

Chapter 7

Conclusion and perspectives

MAS has gone from strength to strength over the last few years and has become a precise and accurate alternative to manual segmentation. By using prior information from manually delineated images (atlases), MAS approaches are capable of capturing the large variability of anatomical structures. However, most MAS methods still suffer from high computational costs at runtime, mainly due to (a) the many nonlinear registrations they require to estimate the spatial correspondence between atlases and target and/or (b) the complexity of label fusion. In this thesis, we have presented a new approach that drastically reduces computational costs at runtime while still providing state-of-the-art segmentation accuracy.

7.1 Chapter overview

Chapter 1 (Introduction) outlined the importance of consistent and accurate segmentation of anatomical structures for diagnosis, monitoring of disease progression or treatment planning, which lead to the development of our approach. Techniques that incorporate a high level of prior information for the segmentation of complex anatomical structures with large variability in shape and appearance, such as the brain, have been of particular interest. MAS has become one of the most successful strategies by leveraging prior

information from multiple individual atlases. The segmentation accuracy of a MAS algorithm is mainly governed by the quality of the registration between each atlas and the target, and the precision of the manual atlas labels. The computational efficiency at runtime is mainly influenced by the number and type of registrations. Considering both accuracy and efficiency for each component of the MAS algorithm, we approached these challenges by shifting time-consuming tasks such as the creation of a population template and learning from the atlases offline, leaving only two registrations online. At runtime we measure similarity locally between the atlas images and the target in manifold space and use the measures to efficiently construct a target-specific template (TST) which further improves registration accuracy. The similarity measures are also applied in the local fusion of the candidate labels.

Chapter 2 (Anatomical atlas building and multi-atlas segmentation) introduced different solutions for estimating the spatial correspondence between all atlas and target images. With an interest in reducing the computational burden at runtime, we focused on the creation of a population-based (average) template from the atlases offline, which can subsequently be used as a space for normalisation for the target. Several pre-processing tasks such as intensity inhomogeneity correction, skull-stripping and tissue classification are usually applied before creating this template. The potential impact of the accuracy, with which every step is performed, on the quality of subsequent steps stresses the necessity for an accurate pre-processing pipeline in our approach. Another important factor which influences the quality and computational efficiency of the template creation process and the estimation of deformation fields, is the type of registration. Since the subsequent steps of our algorithm made inferences from the deformation fields, both accuracy and speed were important factors when comparing several pre-processing and registration methods in our experiments. We noticed large differences in the quality of the results when comparing skull-stripping methods, which is further complicated by our use of different datasets. Freesurfer, which

slightly under-segmented the brain, was chosen, since it might introduce less errors in subsequent steps than methods that over-segment and remove parts of brain tissue. The comparison of two image registration and template creation methods showed crisper templates and smoother deformation fields with SyN compared to DARTEL. However, DARTEL was chosen for all subsequent registrations, since it was computationally more efficient in our tests. In conclusion, results from the pre-processing pipeline require careful examination after every step to ensure high quality. The use of population templates allows computationally more efficient estimation of the spatial correspondence of atlases and targets. Segmentation accuracy is influenced by the similarity and registration accuracy between target and template, which can be refined with a target-specific template (TST).

Chapter 3 (Target-specific template) concerned the use of intermediary images, selected from the atlas set or constructed as a template thereof, to improve registration accuracy between atlases and a target. We described similarity measurement methods for the selection process and their characteristics in utilising various bases and metrics to determine a similarity ranking between the atlases and the target. We presented our approach to constructing a TST of high similarity to the target, which served as an intermediary registration step to enhance registration quality between the target and population template. The TST was constructed from the locally most similar atlases to the target. Similarity was measured between corresponding ROIs of the atlas- and target deformation fields to further minimise the impact of artefacts or intensity inhomogeneities, often present in intensity images. The measurement was performed in low-dimensional manifold spaces, which considered the intrinsic structure of the data leading to more accurate distance measures between atlas- and target projections and, consequently to more accurate weights for TST construction and label fusion detailed in the next chapter. After comparing manifold embedding strategies Isomap was chosen for subsequent steps, since it showed similar accuracy to LLE. Linear

methods such as PCA, used in the eigenimage approach, showed to be less suitable for the reconstruction of target images which could be caused by the high variability in MR brain images. Due to the high similarity of the TST to the target, registration errors could be reduced. Consistent accuracy improvement was achieved on all datasets when our TST was used, showing its robustness to different label maps and image contrasts.

Chapter 4 (Label fusion) outlined strategies for the fusion of selected candidate labels in the space of the target. We described several established methods and the use of our manifold distances as weights for the use in locally weighted fusion. Estimating the intensity distribution from high-probability regions allowed further refinement to detect the structure as a whole. We also presented a hierarchical solution which, in addition to the weights from the manifold, considered the similarity between atlases and target in local patches around anatomical structures of high variability. In our experiments the intensity-based thresholding approach showed robust results for four out of six datasets. However, it required the adjustment of parameters to the dataset and was inferior to fixed thresholding methods in datasets with poor contrast. Weighted fusion achieved the most accurate results on datasets with a relatively small number of atlases. In datasets with a large number of atlases unweighted fusion reached a similar level of accuracy to weighted fusion. To further improve weighted fusion in large datasets the weighting function could be adjusted. Our local patch-based refinement showed significant improvement, while being computationally efficient compared to other patch-based methods. Overall, our fusion approach in combination with the TST, reached state-of-the-art accuracy for all datasets while being computationally much more efficient.

Chapter 5 (Dynamically adjusted labels) presented an overview of methods to further learn or extract information from the atlas dataset offline. The information can then be utilised at runtime without additional

computational cost. One specific area of interest was to determine sub-ROIs from the labels in the atlas dataset, which can have a positive impact on accuracy and reduce the number of candidates for fusion. Confronted with goal of finding meaningful sub-ROIs from the atlas labels, we presented a novel solution by clustering the ROIs of the deformation fields. The resulting clusters indicated regions of high variability, where anatomical structures are less well preserved in the atlas population, and low variability, where structures are more preserved. The clusters were used for TST construction, and candidate selection and -fusion. The recombination of the clusters back into dataset-specific labels was incorporated into our fusion method based on locally weighted fusion. Depending on the variant of the used fusion strategy, the dataset, and the number of clusters, our method showed mixed results. We believe the selection of the number of clusters is an important factor and depends on ROI size and anatomy. This issue becomes more challenging with more diverse ROIs and requires more tests in the future. Overall our approach holds potential for improving segmentation accuracy and/or reducing the number of candidates for fusion by estimating the ideal number of candidates for each cluster. The estimation could incorporate prior knowledge such as the level of preservation in a ROI.

Chapter 6 (Application to Tourette's images) presented a test case for the initial aim of our method: to locate and segment anatomical structures and allow comparison of healthy and diseased populations. We applied our approach, explained in Chapters 2-4, to MR brain scans of adolescent healthy subjects and Tourette's patients. We segmented and compared the volumes of subcortical structures, which were of particular interest in our study. We found a significantly different growth pattern in the right pallidum in TS subjects compared to HC. Due to the focus of the study on an adolescent population, our findings showed similarities to results from both children and adults in the literature, providing complementary information for the age gap in between.

7.2 Conclusion

Two challenges in medical image segmentation can be defined as the identification of image features and their translation into semantic entities to provide context. Depending on the variability in shape and appearance of an anatomical structure, the success of a method often depends on the incorporated level of prior knowledge. Multi-atlas segmentation (MAS) strategies utilise both explicit knowledge from human operators in the form of manual delineations of anatomical structures of interest and implicit knowledge from image intensities or parameters derived thereof. Motivated by the necessity to segment MR brain images both accurately and in a time efficient manner, we presented a novel MAS method with the following advantages.

The use of an average template as a space for normalisation reduced the number of nonlinear registrations at runtime to two, between the target and the average population template, and between the target and the TST. This is in stark contrast to other top performing methods which require as many nonlinear registrations as there are atlases at runtime. Special consideration should be given to the pre-processing methods, since in our experiment large differences in quality and robustness were found.

Registering the target to a highly similar TST reduced the registration error by dividing the task of estimating a large difficult deformation into smaller more precise deformations. Higher accuracy was achieved in our test when using the TST as intermediate step on all datasets. The composition of deformation fields for the propagation of label maps showed no reduction in segmentation accuracy as observed for other methods in the literature. This could be due to our use of a high-dimensional registration method and the use of the TST for improving registration accuracy.

The use of nonlinear manifold embedding strategies for the selection of the most similar atlases to the target had the advantage of finding the intrinsic structure in the data and allowed more accurate similarity and weight estimation, derived from the distance measurements in low-dimensional space. The

use of deformation fields as basis for comparison has the advantage of being less influenced by artefacts or intensity inhomogeneities. These weights were used for the construction of the TST and the fusion of the candidate labels, which showed improved segmentation accuracy for small datasets. For large datasets the unweighted fusion variant showed similar results, since more suitable candidates were available. Adjustments to the weighting function are expected to further increase accuracy with the weighted fusion variant. Our hierarchical approach with patch-refinement showed further improvements by providing even more localised weights in lower-probability regions, while being computationally more efficient than other patch-based methods.

The clustering of regions of interest (ROIs) into smaller, more regional sub-ROIs allowed more accurate weight estimation for TST creation and label fusion in some datasets. In other datasets it had a negative impact which could be caused by the dataset-specific label map, making the choice of cluster numbers challenging. Considering that our MAS algorithm could potentially be used with a large atlas dataset, shifting the computationally expensive tasks offline becomes more important.

Overall, our method showed robust, state-of-the-art segmentation accuracy in several datasets with only minimal parameter-tuning. Compared to other top-performing methods, our approach can drastically reduce the computational costs at runtime.

7.3 Perspectives

Larger atlas datasets from diverse populations: Since the computational efficiency of our method does not depend on the number of atlases, it is scalable and would potentially allow the use of large atlas datasets from diverse populations. In a larger dataset the chance of finding a similar atlas or a part thereof to a target could be increased and potentially further improve registration and fusion accuracy.

Label fusion: We plan to test our method with more sophisticated fusion strategies such as joint label fusion [221] and also develop our own method to make better use of the information from clusters. Although a weighted and unweighted fusion strategy for recombination of the clusters into the dataset-specific ROIs was presented, further improvements are required to construct smooth borders between neighbouring clusters.

Clustering: In the future we will test clustering methods that do not require a pre-defined choice of cluster numbers, e.g. affinity propagation [83]. Clustering the ROIs based on their level of anatomical preservation has the potential to allow the selection of an ideal number of candidates dynamically for each sub-ROI. The fusion in more preserved sub-ROIs of the dataset could use more candidates due to their high similarity to the target. Conversely, for less preserved sub-ROIs a smaller number of candidates might be beneficial to introduce less errors by unsuitable lower-ranked candidates. We are also working on a theoretically sound framework for the selection of the optimal number of atlases. In addition to location and magnitude of the deformation fields we would like to employ more spatial information for finding clusters.

Anatomy: Since the label maps of brain atlases usually include relatively small labels, we would also like to test our approach on larger anatomical structures such as the heart.

Bibliography

[1] O. Acosta, A. Simon, F. Monge, F. Commandeur, C. Bassirou, G. Cazoulat, R. de Crevoisier, and P. Haigron. Evaluation of multi-atlas-based segmentation of CT scans in prostate cancer radiotherapy. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1966–1969, March 2011.

[2] R. Adams and L. Bischof. Seeded region growing. *IEEE Transactions on pattern analysis and machine intelligence*, 16(6):641–647, 1994.

[3] S. Alchatzidis, A. Sotiras, and N. Paragios. Discrete multi atlas segmentation using agreement constraints. In *Proceedings of the British Machine Vision Conference*. BMVA Press, 2014.

[4] 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.

[5] P. Aljabar, R. Heckemann, A. Hammers, J. V. Hajnal, and D. Rueckert. Classifier selection strategies for label fusion using large atlas databases. In N. Ayache, S. Ourselin, and A. Maeder, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2007: 10th International Conference, Brisbane, Australia, October 29 - November 2, 2007, Proceedings, Part I*, pages 523–531, Berlin, Heidelberg, 2007. Springer Berlin Heidelberg.

[6] P. Aljabar, R. Wolz, and D. Rueckert. Manifold learning for medical

image registration, segmentation, and classification. *Machine Learning in Computer-Aided Diagnosis: Medical Imaging Intelligence and Analysis: Medical Imaging Intelligence and Analysis*, page 351, 2012.

[7] J. S. Allen, H. Damasio, and T. J. Grabowski. Normal neuroanatomical variation in the human brain: An MRI-volumetric study. *American Journal of Physical Anthropology*, 118(4):341–358, 2002.

[8] J. S. Allen, H. Damasio, T. J. Grabowski, J. Bruss, and W. Zhang. Sexual dimorphism and asymmetries in the graywhite composition of the human cerebrum. *NeuroImage*, 18(4):880 – 894, 2003.

[9] K. Amunts, A. Malikovic, H. Mohlberg, T. Schormann, and K. Zilles. Brodmann's areas 17 and 18 brought into stereotaxic space - where and how variable? *NeuroImage*, 11(1):66 – 84, 2000.

[10] B. Aribisala, S. Cox, K. Ferguson, S. MacPherson, A. MacLullich, N. Royle, M. V. Hernández, M. Bastin, I. Deary, and J. Wardlaw. Assessing the performance of atlas-based prefrontal brain parcellation in an ageing cohort. *Journal of computer assisted tomography*, 37(2), 2013.

[11] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de Solorzano. Combination strategies in multi-atlas image segmentation: Application to brain MR data. *Medical Imaging, IEEE Transactions on*, 28(8):1266–1277, Aug 2009.

[12] X. Artaechevarria, A. Munoz-Barrutia, and C. Ortiz-de Solrzano. Efficient classifier generation and weighted voting for atlas-based segmentation: two small steps faster and closer to the combination oracle. *Proc.SPIE*, 6914:6914 – 6914 – 9, 2008.

[13] C. Arthofer, P. S. Morgan, and A. Pitiot. Hierarchical multi-atlas segmentation using label-specific embeddings, target-specific templates and patch refinement. In G. Wu, P. Coupé, Y. Zhan, B. C. Munsell,

and D. Rueckert, editors, *Patch-Based Techniques in Medical Imaging: Second International Workshop, Patch-MI 2016, Held in Conjunction with MICCAI 2016, Athens, Greece, October 17, 2016, Proceedings*, pages 84–91, Cham, 2016. Springer International Publishing.

[14] J. Ashburner. A fast diffeomorphic image registration algorithm. *NeuroImage*, 38(1):95 – 113, 2007.

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

[16] J. Ashburner and K. J. Friston. Computing average shaped tissue probability templates. *NeuroImage*, 45(2):333 – 341, 2009.

[17] A. J. Asman, A. Akhondi-Asl, H. Wang, N. Tustison, B. Avants, S. K. Warfield, and B. Landman. MICCAI 2013 segmentation algorithms, theory and applications (SATA) challenge results summary. `https://my.vanderbilt.edu/masi/workshops/`, 2013. Accessed: 16/09/2017.

[18] A. J. Asman and B. A. Landman. Formulating spatially varying performance in the statistical fusion framework. *IEEE Transactions on Medical Imaging*, 31(6):1326–1336, June 2012.

[19] A. J. Asman and B. A. Landman. Multi-atlas segmentation using non-local STAPLE. In *MICCAI Workshop on Multi-Atlas Labeling*, pages 87–90, 2012.

[20] A. J. Asman and B. A. Landman. Non-local statistical label fusion for multi-atlas segmentation. *Medical Image Analysis*, 17(2):194 – 208, 2013.

[21] B. B. Avants, C. L. Epstein, M. Grossman, and J. C. Gee. Symmetric diffeomorphic image registration with cross-correlation: Evaluating automated labeling of elderly and neurodegenerative brain. *Medical Image Analysis*, 12(1):26–41, 09 2007.

[22] B. B. Avants, N. J. Tustison, J. Wu, P. A. Cook, and J. C. Gee. An open source multivariate framework for n-tissue segmentation with evaluation on public data. *Neuroinformatics*, 9(4):381–400, Dec 2011.

[23] B. B. Avants, P. Yushkevich, J. Pluta, D. Minkoff, M. Korczykowski, J. Detre, and J. C. Gee. The optimal template effect in hippocampus studies of diseased populations. *NeuroImage*, 49(3):2457 – 2466, 2010.

[24] W. F. Baar, H. E. Hulshoff Pol, D. I. Boomsma, D. Posthuma, E. J. de Geus, H. G. Schnack, N. E. van Haren, C. J. van Oel, and R. S. Kahn. Quantitative genetic modeling of variation in human brain morphology. *Cerebral Cortex*, 11(9):816–824, 2001.

[25] W. Bai, W. Shi, D. P. O'Regan, T. Tong, H. Wang, S. Jamil-Copley, N. S. Peters, and D. Rueckert. A probabilistic patch-based label fusion model for multi-atlas segmentation with registration refinement: Application to cardiac MR images. *IEEE Transactions on Medical Imaging*, 32(7):1302–1315, July 2013.

[26] R. Bajcsy and S. Kovai. Multiresolution elastic matching. *Computer Vision, Graphics, and Image Processing*, 46(1):1 – 21, 1989.

[27] 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(4), 1983.

[28] I. N. Bankman. Segmentation. In I. N. BANKMAN, editor, *Handbook of Medical Image Processing and Analysis (Second Edition)*, pages 71 – 72. Academic Press, Burlington, second edition edition, 2009.

[29] M. F. Beg, M. I. Miller, A. Trouvé, and L. Younes. Computing large deformation metric mappings via geodesic flows of diffeomorphisms. *International Journal of Computer Vision*, 61(2):139–157, Feb 2005.

[30] O. M. Benkarim, G. Piella, M. A. G. Ballester, and G. Sanroma. Discriminative confidence estimation for probabilistic multi-atlas label fusion. *Medical Image Analysis*, 42:274 – 287, 2017.

[31] O. M. Benkarim, G. Piella, M. A. González Ballester, and G. Sanroma. Enhanced probabilistic label fusion by estimating label confidences through discriminative learning. In S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part II*, pages 505–512, Cham, 2016. Springer International Publishing.

[32] S. Beucher and C. Lantujoul. Use of watersheds in contour detection, Sept. 1979.

[33] K. K. Bhatia, J. V. Hajnal, B. K. Puri, A. D. Edwards, and D. Rueckert. Consistent groupwise non-rigid registration for atlas construction. In *2004 2nd IEEE International Symposium on Biomedical Imaging: Nano to Macro (IEEE Cat No. 04EX821)*, pages 908–911 Vol. 1, April 2004.

[34] M. H. Bloch, J. F. Leckman, H. Zhu, and B. S. Peterson. Caudate volumes in childhood predict symptom severity in adults with tourette syndrome. *Neurology*, 65(8):1253–1258, 2005.

[35] M. Boccardi, M. Bocchetta, F. C. Morency, D. L. Collins, M. Nishikawa, R. Ganzola, M. J. Grothe, D. Wolf, A. Redolfi, M. Pievani, L. Antelmi, A. Fellgiebel, H. Matsuda, S. Teipel, S. Duchesne, J. Jack, Clifford R., and G. B. Frisoni. Training labels for hippocampal segmentation based on the EADC-ADNI harmonized hippocampal protocol. *Alzheimer's & Dementia: The Journal of the Alzheimer's Association*, 11(2):175–183, 08 2014.

[36] J. A. Bogovic, B. Jedynak, R. Rigg, A. Du, B. A. Landman, J. L. Prince, and S. H. Ying. Approaching expert results using a hierarchical cerebellum parcellation protocol for multiple inexpert human raters. *NeuroImage*, 64(Supplement C):616 – 629, 2013.

[37] F. L. Bookstein. Principal warps: thin-plate splines and the decomposition of deformations. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 11(6):567–585, Jun 1989.

[38] L. Breiman. Bagging predictors. *Machine Learning*, 24(2):123–140, Aug 1996.

[39] C. R. Brice and C. L. Fennema. Scene analysis using regions. *Artificial Intelligence*, 1(3):205 – 226, 1970.

[40] F. W. Bryan, Z. Xu, A. J. Asman, W. M. Allen, D. S. Reich, and B. A. Landman. Self-assessed performance improves statistical fusion of image labels. *Medical Physics*, 41(3):031903, 2014.

[41] A. Buades, B. Coll, and J.-M. Morel. A non-local algorithm for image denoising. In *Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2 - Volume 02*, CVPR '05, pages 60–65, Washington, DC, USA, 2005. IEEE Computer Society.

[42] J. Canny. A computational approach to edge detection. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-8(6):679–698, Nov 1986.

[43] Y. Cao, X. Li, and P. Yan. Multi-atlas based image selection with label image constraint. In *2012 11th International Conference on Machine Learning and Applications*, volume 1, pages 311–316, Dec 2012.

[44] Y. Cao, Y. Yuan, X. Li, B. Turkbey, P. Choyke, and P. Yan. Segmenting images by combining selected atlases on manifold. In G. Fichtinger,

A. Martel, and T. Peters, editors, *Medical Image Computing and Computer-Assisted Intervention MICCAI 2011*, volume 6893 of *Lecture Notes in Computer Science*, pages 272–279. Springer Berlin Heidelberg, 2011.

[45] Y. Cao, Y. Yuan, X. Li, and P. Yan. Putting images on a manifold for atlas-based image segmentation. In *2011 18th IEEE International Conference on Image Processing*, pages 289–292, Sept 2011.

[46] M. Cardoso, M. Modat, R. Wolz, A. Melbourne, D. Cash, D. Rueckert, and S. Ourselin. Geodesic information flows: Spatially-variant graphs and their application to segmentation and fusion. *Medical Imaging, IEEE Transactions on*, 34(9):1976–1988, Sept 2015.

[47] M. J. Cardoso, K. Leung, M. Modat, S. Keihaninejad, D. Cash, J. Barnes, N. C. Fox, and S. Ourselin. STEPS: Similarity and truth estimation for propagated segmentations and its application to hippocampal segmentation and brain parcelation. *Medical Image Analysis*, 17(6):671 – 684, 2013.

[48] K. R. Castleman. *Digital Image Processing*. Prentice Hall Press, Upper Saddle River, NJ, USA, 1996.

[49] G. E. Christensen, X. Geng, J. G. Kuhl, J. Bruss, T. J. Grabowski, I. A. Pirwani, M. W. Vannier, J. S. Allen, and H. Damasio. Introduction to the non-rigid image registration evaluation project (NIREP). In J. P. W. Pluim, B. Likar, and F. A. Gerritsen, editors, *Third International Workshop, WBIR 2006, Proceedings*, pages 128–135. Springer Berlin Heidelberg, 2006.

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

[51] G. E. Christensen, R. D. Rabbitt, and M. I. Miller. Deformable templates using large deformation kinematics. *IEEE Transactions on Image Processing*, 5(10):1435–1447, Oct 1996.

[52] J. A. Church and B. L. Schlaggar. Pediatric Tourette syndrome: Insights from recent neuroimaging studies. *Journal of Obsessive-Compulsive and Related Disorders*, 3(4):386 – 393, 2014.

[53] L. Clarke, R. Velthuizen, M. Camacho, J. Heine, M. Vaidyanathan, L. Hall, R. Thatcher, and M. Silbiger. MRI segmentation: Methods and applications. *Magnetic Resonance Imaging*, 13(3):343 – 368, 1995.

[54] L. D. Cohen. On active contour models and balloons. *CVGIP: Image Understanding*, 53(2):211 – 218, 1991.

[55] A. Collignon, F. Maes, D. Delaere, D. Vandermeulen, P. Suetens, and G. Marchal. Automated multi-modality image registration based on information theory. In *Lect Notes Comput Sci*, volume 3, pages 263–274, 01 1995.

[56] 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.

[57] D. L. Collins, P. Neelin, T. M. Peters, and A. C. Evans. Automatic 3D intersubject registration of MR volumetric data in standardized Talairach space. *J Comput Assist Tomogr*, 18(2):192–205, Mar-Apr 1994.

[58] D. L. Collins, A. P. Zijdenbos, V. Kollokian, J. G. Sled, N. J. Kabani, C. J. Holmes, and A. C. Evans. Design and construction of a realistic digital brain phantom. *IEEE Transactions on Medical Imaging*, 17(3):463–468, June 1998.

[59] O. Commowick, V. Grégoire, and G. Malandain. Atlas-based delineation of lymph node levels in head and neck computed tomography images. *Radiotherapy and Oncology*, 87(2):281–289, 2008.

[60] O. Commowick and G. Malandain. Efficient selection of the most similar image in a database for critical structures segmentation. In N. Ayache, S. Ourselin, and A. Maeder, editors, *Medical Image Computing and Computer-Assisted Intervention MICCAI 2007*, volume 4792 of *Lecture Notes in Computer Science*, pages 203–210. Springer Berlin Heidelberg, 2007.

[61] O. Commowick, S. K. Warfield, and G. Malandain. Using Frankenstein's creature paradigm to build a patient specific atlas. In G.-Z. Yang, D. Hawkes, D. Rueckert, A. Noble, and C. Taylor, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2009: 12th International Conference, London, UK, September 20-24, 2009, Proceedings, Part II*, pages 993–1000, Berlin, Heidelberg, 2009. Springer Berlin Heidelber.

[62] T. F. Cootes, G. J. Edwards, and C. J. Taylor. Active appearance models. In H. Burkhardt and B. Neumann, editors, *Computer Vision — ECCV'98: 5th European Conference on Computer Vision Freiburg, Germany, June 2–6, 1998 Proceedings, Volume II*, pages 484–498. Springer Berlin Heidelberg, Berlin, Heidelberg, 1998.

[63] T. F. Cootes, C. J. Taylor, D. H. Cooper, and J. Graham. Active shape models-their training and application. *Computer vision and image understanding*, 61(1):38–59, 1995.

[64] P. Coupé, J. V. Manjón, V. Fonov, J. Pruessner, M. Robles, and D. L. Collins. Patch-based segmentation using expert priors: Application to hippocampus and ventricle segmentation. *NeuroImage*, 54(2):940 – 954, 2011.

[65] W. Crum, L. Griffin, D. Hill, and D. Hawkes. Zen and the art of medical image registration: correspondence, homology, and quality. *NeuroImage*, 20(3):1425 – 1437, 2003.

[66] W. R. Crum, O. Camara, and D. L. G. Hill. Generalized overlap measures for evaluation and validation in medical image analysis. *IEEE Transactions on Medical Imaging*, 25(11):1451–1461, Nov 2006.

[67] A. M. Dale, B. Fischl, and M. I. Sereno. Cortical surface-based analysis: I. Segmentation and surface reconstruction. *NeuroImage*, 9(2):179 – 194, 1999.

[68] C. Davatzikos. Spatial normalization of 3D brain images using deformable models. *J Comput Assist Tomogr*, 20(4):656–665, Jul-Aug 1996.

[69] B. M. Dawant, S. L. Hartmann, J. P. Thirion, F. Maes, D. Vandermeulen, and P. Demaerel. Automatic 3-D segmentation of internal structures of the head in MR images using a combination of similarity and free-form transformations. i. Methodology and validation on normal subjects. *IEEE Transactions on Medical Imaging*, 18(10):909–916, Oct 1999.

[70] M. Depa, G. Holmvang, E. J. Schmidt, P. Golland, and M. R. Sabuncu. Towards effcient label fusion by pre-alignment of training data. In *Medical image computing and computer-assisted intervention: MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention*, volume 14, page 38. NIH Public Access, 2011.

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

[72] N. T. Doan, J. O. de Xivry, and B. Macq. Effect of inter-subject variation on the accuracy of atlas-based segmentation applied to human brain structures. In B. M. Dawant and D. R. Haynor, editors, *Medical*

Imaging: Image Processing, volume 7623 of *SPIE Proceedings*, page 76231S. SPIE, 2010.

[73] J. Doshi, G. Erus, Y. Ou, B. Gaonkar, and C. Davatzikos. Multi-atlas skull-stripping. *Academic Radiology*, 20(12):1566–1576, 2013.

[74] J. Doshi, G. Erus, Y. Ou, S. M. Resnick, R. C. Gur, R. E. Gur, T. D. Satterthwaite, S. Furth, and C. Davatzikos. MUSE: Multi-atlas region segmentation utilizing ensembles of registration algorithms and parameters, and locally optimal atlas selection. *NeuroImage*, 127:186 – 195, 2016.

[75] A. K. H. Duc, M. Modat, K. K. Leung, M. J. Cardoso, J. Barnes, T. Kadir, and S. Ourselin. Using manifold learning for atlas selection in multi-atlas segmentation. *PLoS ONE*, 8(8):e70059, 08 2013.

[76] A. Elnakib, G. Gimel'farb, J. S. Suri, and A. El-Baz. Medical image segmentation: A brief survey. In A. S. El-Baz, R. Acharya U, A. F. Laine, and J. S. Suri, editors, *Multi Modality State-of-the-Art Medical Image Segmentation and Registration Methodologies: Volume II*, pages 1–39, New York, NY, 2011. Springer New York.

[77] S. F. Eskildsen, P. Coup, V. Fonov, J. V. Manjn, K. K. Leung, N. Guizard, S. N. Wassef, L. R. stergaard, and D. L. Collins. BEaST: Brain extraction based on nonlocal segmentation technique. *NeuroImage*, 59(3):2362 – 2373, 2012.

[78] A. Evans, D. Collins, S. Mills, E. Brown, R. Kelly, and T. Peters. 3D statistical neuroanatomical models from 305 MRI volumes. In *Nuclear Science Symposium and Medical Imaging Conference, 1993., 1993 IEEE Conference Record.*, pages 1813–1817 vol.3, Oct 1993.

[79] W. Feng, S. J. Reeves, T. S. Denney, S. Lloyd, L. Dell'Italia, and H. Gupta. A new consistent image registration formulation with a B-

spline deformation model. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 979–982, June 2009.

[80] 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.

[81] V. Fonov, P. Coupé, S. F. Eskildsen, J. V. Manjon, and L. Collins. Multi-atlas labeling with population-specific template and non-local patch-based label fusion. In *MICCAI 2012 Workshop on Multi-Atlas Labeling*, pages 63–66, Nice, France, Oct. 2012.

[82] E. Forgy. Cluster analysis of multivariate data: Efficiency versus interpretability of classification. *Biometrics*, 21(3):768–769, 1965.

[83] B. J. Frey and D. Dueck. Clustering by passing messages between data points. *Science*, 315(5814):972–976, 2007.

[84] G. B. Frisoni, C. R. Jack, M. Bocchetta, C. Bauer, K. S. Frederiksen, Y. Liu, G. Preboske, T. Swihart, M. Blair, E. Cavedo, M. J. Grothe, M. Lanfredi, O. Martinez, M. Nishikawa, M. Portegies, T. Stoub, C. Ward, L. G. Apostolova, R. Ganzola, D. Wolf, F. Barkhof, G. Bartzokis, C. DeCarli, J. G. Csernansky, L. deToledo Morrell, M. I. Geerlings, J. Kaye, R. J. Killiany, S. Lehéricy, H. Matsuda, J. O'Brien, L. C. Silbert, P. Scheltens, H. Soininen, S. Teipel, G. Waldemar, A. Fellgiebel, J. Barnes, M. Firbank, L. Gerritsen, W. Henneman, N. Malykhin, J. C. Pruessner, L. Wang, C. Watson, H. Wolf, M. deLeon, J. Pantel, C. Ferrari, P. Bosco, P. Pasqualetti, S. Duchesne, H. Duvernoy, and M. Boccardi. The EADC-ADNI harmonized protocol for manual hippocampal segmentation on magnetic resonance: Evidence of validity. *Alzheimer's & dementia : the journal of the Alzheimer's Association*, 11(2):111–125, 02 2015.

[85] K. J. Friston, J. Ashburner, C. D. Frith, J.-B. Poline, J. D. Heather, and R. S. J. Frackowiak. Spatial registration and normalization of images. *Human Brain Mapping*, 3(3):165–189, 1995.

[86] M. Ganz, D. Kondermann, J. Andrulis, G. M. Knudsen, and L. Maier-Hein. Crowdsourcing for error detection in cortical surface delineations. *International Journal of Computer Assisted Radiology and Surgery*, 12(1):161–166, Jan 2017.

[87] M. Ganzetti, N. Wenderoth, and D. Mantini. Quantitative evaluation of intensity inhomogeneity correction methods for structural MR brain images. *Neuroinformatics*, 14(1):5–21, Jan 2016.

[88] Q. Gao, T. Tong, D. Rueckert, and P. Edwards. Multi-atlas propagation via a manifold graph on a database of both labeled and unlabeled images. *Proc. SPIE*, 9035:90350A–90350A–7, 2014.

[89] S. Gerber, T. Tasdizen, S. Joshi, and R. Whitaker. On the manifold structure of the space of brain images. *Medical image computing and computer-assisted intervention : MICCAI International Conference on Medical Image Computing and Computer-Assisted Intervention*, 12(0 1):305–312, 2009.

[90] A. Gholipour, C. K. Rollins, C. Velasco-Annis, A. Ouaalam, A. Akhondi-Asl, O. Afacan, C. M. Ortinau, S. Clancy, C. Limperopoulos, E. Yang, J. A. Estroff, and S. K. Warfield. A normative spatiotemporal MRI atlas of the fetal brain for automatic segmentation and analysis of early brain growth. *Scientific Reports*, 7(1):476, 2017.

[91] R. C. Gonzalez and R. E. Woods. *Digital Image Processing (3rd Edition)*. Prentice-Hall, Inc., Upper Saddle River, NJ, USA, 2006.

[92] I. S. Gousias, A. D. Edwards, M. A. Rutherford, S. J. Counsell, J. V. Hajnal, D. Rueckert, and A. Hammers. Magnetic resonance imaging of

the newborn brain: Manual segmentation of labelled atlases in term-born and preterm infants. *NeuroImage*, 62(3):1499 – 1509, 2012.

[93] P. Grandjean and P. Landrigan. Developmental neurotoxicity of industrial chemicals. *The Lancet*, 368(9553):2167 – 2178, 2006.

[94] T. Greitz, C. Bohm, S. Holte, and L. Eriksson. A computerized brain atlas: Construction, anatomical content, and some applications. *Journal of Computer Assisted Tomography*, 15(1), 1991.

[95] A. Guimond, J. Meunier, and J.-P. Thirion. Average brain models: A convergence study. *Computer Vision and Image Understanding*, 77(2):192 – 210, 2000.

[96] 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. Special Issue on the 12th International Conference on Medical Image Computing and Computer-Assisted Intervention (MICCAI) 2009.

[97] Y. Hao, T. Jiang, and Y. Fan. Iterative multi-atlas based segmentation with multi-channel image registration and Jackknife context model. In *2012 9th IEEE International Symposium on Biomedical Imaging (ISBI)*, pages 900–903, May 2012.

[98] 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.

[99] R. A. Heckemann, J. V. Hajnal, P. Aljabar, D. Rueckert, and A. Hammers. Multiclassifier fusion in human brain MR segmentation: Modelling convergence. In R. Larsen, M. Nielsen, and J. Sporring, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006: 9th International Conference, Copenhagen, Denmark,*

October 1-6, 2006. Proceedings, Part II, pages 815–822, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[100] R. A. Heckemann, A. Hammers, P. Aljabar, D. Rueckert, and J. V. Hajnal. The mirror method of assessing segmentation quality in atlas label propagation. In *2009 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1194–1197, June 2009.

[101] M. P. Heinrich, M. Jenkinson, M. Bhushan, T. Matin, F. V. Gleeson, M. Brady, and J. A. Schnabel. MIND: Modality independent neighbourhood descriptor for multi-modal deformable registration. *Medical Image Analysis*, 16(7):1423–1435, 2012.

[102] C. J. Holmes, R. Hoge, L. Collins, R. Woods, A. W. Toga, and A. C. Evans. Enhancement of MR images using registration for signal averaging. *J Comput Assist Tomogr*, 22(2):324–333, Mar-Apr 1998.

[103] S. L. Horowitz and T. Pavlidis. Picture Segmentation by a directed split-and-merge procedure. *Proceedings of the 2nd International Joint Conference on Pattern Recognition, Copenhagen, Denmark*, pages 424–433, 1974.

[104] J. E. Iglesias, C. Y. Liu, P. M. Thompson, and Z. Tu. Robust brain extraction across datasets and comparison with publicly available methods. *IEEE Transactions on Medical Imaging*, 30(9):1617–1634, Sept 2011.

[105] J. E. Iglesias and M. R. Sabuncu. Multi-atlas segmentation of biomedical images: A survey. *Medical Image Analysis*, 24(1):205 – 219, 2015.

[106] J. E. Iglesias, M. R. Sabuncu, and K. V. Leemput. A unified framework for cross-modality multi-atlas segmentation of brain MRI. *Medical Image Analysis*, 17(8):1181 – 1191, 2013.

[107] I. Isgum, M. Staring, A. Rutten, M. Prokop, M. A. Viergever, and B. van Ginneken. Multi-atlas-based segmentation with local decision fusion - application to cardiac and aortic segmentation in CT scans. *IEEE Transactions on Medical Imaging*, 28(7):1000–1010, July 2009.

[108] P. Jaccard. The distribution of the flora in the alpine zone.1. *New Phytologist*, 11(2):37–50, 1912.

[109] C. R. Jack, 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.

[110] H. Jia, P.-T. Yap, and D. Shen. Iterative multi-atlas-based multi-image segmentation with tree-based registration. *NeuroImage*, 59(1):422 – 430, 2012.

[111] H. Jia, P.-T. Yap, G. Wu, Q. Wang, and D. Shen. Intermediate templates guided groupwise registration of diffusion tensor images. *NeuroImage*, 54(2):928 – 939, 2011.

[112] I. T. Jolliffe. *Principal Component Analysis and Factor Analysis*, pages 115–128. Springer New York, New York, NY, 1986.

[113] S. Joshi, B. Davis, M. Jomier, and G. Gerig. Unbiased diffeomorphic atlas construction for computational anatomy. *NeuroImage*, 23:S151 – S160, 2004. Mathematics in Brain Imaging.

[114] K. Kamnitsas, C. Ledig, V. F. Newcombe, J. P. Simpson, A. D. Kane, D. K. Menon, D. Rueckert, and B. Glocker. Efficient multi-scale 3D

CNN with fully connected CRF for accurate brain lesion segmentation. *Medical Image Analysis*, 36:61 – 78, 2017.

[115] M. Kass, A. Witkin, and D. Terzopoulos. Snakes: Active contour models. *International journal of computer vision*, 1(4):321–331, 1988.

[116] H. Kekre and S. Gharge. Texture based segmentation using statistical properties for mammographic images, 11 2010.

[117] R. Kikinis, M. E. Shenton, D. V. Iosifescu, R. W. McCarley, P. Saiviroonporn, H. H. Hokama, A. Robatino, D. Metcalf, C. G. Wible, C. M. Portas, R. M. Donnino, and F. A. Jolesz. A digital brain atlas for surgical planning, model-driven segmentation, and teaching. *IEEE Transactions on Visualization and Computer Graphics*, 2(3):232–241, Sep 1996.

[118] M. Kim, G. Wu, P. T. Yap, and D. Shen. A general fast registration framework by learning deformation-appearance correlation. *IEEE Transactions on Image Processing*, 21(4):1823–1833, April 2012.

[119] J. Kittler, M. Hatef, R. P. W. Duin, and J. Matas. On combining classifiers. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(3):226–239, Mar 1998.

[120] J. Kittler and J. Illingworth. Minimum error thresholding. *Pattern recognition*, 19(1):41–47, 1986.

[121] J. Kleesiek, G. Urban, A. Hubert, D. Schwarz, K. Maier-Hein, M. Bendszus, and A. Biller. Deep MRI brain extraction: A 3D convolutional neural network for skull stripping. *NeuroImage*, 129(Supplement C):460 – 469, 2016.

[122] A. Klein, J. Andersson, B. A. Ardekani, J. Ashburner, B. Avants, M.-C. Chiang, G. E. Christensen, D. L. Collins, J. Gee, P. Hellier, J. H. Song, M. Jenkinson, C. Lepage, D. Rueckert, P. Thompson, T. Vercauteren,

R. P. Woods, J. J. Mann, and R. V. Parsey. Evaluation of 14 nonlinear deformation algorithms applied to human brain MRI registration. *NeuroImage*, 46(3):786 – 802, 2009.

[123] A. Klein, B. Mensh, S. Ghosh, J. Tourville, and J. Hirsch. Mindboggle: Automated brain labeling with multiple atlases. *BMC Medical Imaging*, 5(1):7, Oct 2005.

[124] S. Klein, M. Staring, K. Murphy, M. A. Viergever, and J. P. W. Pluim. elastix: A toolbox for intensity-based medical image registration. *IEEE Transactions on Medical Imaging*, 29(1):196–205, Jan 2010.

[125] B. A. Landman and S. K. Warfield, editors. *MICCAI 2012 Workshop on Multi-Atlas Labeling*, 2012.

[126] T. Langerak, U. van der Heide, A. Kotte, F. Berendsen, and J. Pluim. Improving label fusion in multi-atlas based segmentation by locally combining atlas selection and performance estimation. *Computer Vision and Image Understanding*, 130:71 – 79, 2015.

[127] T. R. Langerak, F. F. Berendsen, U. A. Van der Heide, A. N. T. J. Kotte, and J. P. W. Pluim. Multiatlas-based segmentation with preregistration atlas selection. *Medical Physics*, 40(9):091701–n/a, 2013. 091701.

[128] T. R. Langerak, U. A. van der Heide, A. N. T. J. Kotte, F. F. Berendsen, and J. P. W. Pluim. Local atlas selection and performance estimation in multi-atlas based segmentation. In *2011 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 669–672, March 2011.

[129] T. R. Langerak, U. A. van der Heide, A. N. T. J. Kotte, M. A. Viergever, M. van Vulpen, and J. P. W. Pluim. Label fusion in atlas-based segmentation using a selective and iterative method for per-

formance level estimation (SIMPLE). *IEEE Transactions on Medical Imaging*, 29(12):2000–2008, Dec 2010.

[130] C. T. Larsen, J. E. Iglesias, and K. Van Leemput. N3 bias field correction explained as a bayesian modeling method. In M. J. Cardoso, I. Simpson, T. Arbel, D. Precup, and A. Ribbens, editors, *Bayesian and grAphical Models for Biomedical Imaging*, pages 1–12, Cham, 2014. Springer International Publishing.

[131] C. Ledig, R. Wolz, P. Aljabar, J. Ltjnen, R. A. Heckemann, A. Hammers, and D. Rueckert. Multi-class brain segmentation using atlas propagation and EM-based refinement. In *9th IEEE ISBI*, pages 896–899, 2012.

[132] J.-S. Lee, S.-S. Yoo, S.-Y. Cho, S.-M. Ock, M.-K. Lim, and L. P. Panych. Abnormal thalamic volume in treatment-naive boys with Tourette syndrome. *Acta Psychiatrica Scandinavica*, 113(1):64–67, 2006.

[133] R. K. Lenroot and J. N. Giedd. Brain development in children and adolescents: Insights from anatomical magnetic resonance imaging. *Neuroscience & Biobehavioral Reviews*, 30(6):718 – 729, 2006.

[134] B. Li, S. Panda, Z. Xu, A. J. Asman, P. L. Shanahan, R. G. Abramson, and B. A. Landman. Regression forest region recognition enhances multi-atlas spleen labeling. In *MICCAI Challenge Workshop on Segmentation: Algorithms, Theory and Applications (SATA)*, pages 82–92. Citeseer, 2013.

[135] S. Lloyd. Least squares quantization in PCM. *IEEE transactions on information theory*, 28(2):129–137, 1982.

[136] G. Ma, Y. Gao, G. Wu, L. Wu, and D. Shen. Nonlocal atlas-guided multi-channel forest learning for human brain labeling. *Medical Physics*, 43(2):1003–1019, 2016.

[137] J. B. MacQueen. Some methods for classification and analysis of multivariate observations. In L. M. L. Cam and J. Neyman, editors, *Proc. of the fifth Berkeley Symposium on Mathematical Statistics and Probability*, volume 1, pages 281–297. University of California Press, 1967.

[138] F. Maes, A. Collignon, D. Vandermeulen, G. Marchal, and P. Suetens. Multimodality image registration by maximization of mutual information. *IEEE transactions on medical imaging*, 16(2):187–198, 1997.

[139] L. Maier-Hein, S. Mersmann, D. Kondermann, C. Stock, H. G. Kenngott, A. Sanchez, M. Wagner, A. Preukschas, A.-L. Wekerle, S. Helfert, S. Bodenstedt, and S. Speidel. Crowdsourcing for reference correspondence generation in endoscopic images. In P. Golland, N. Hata, C. Barillot, J. Hornegger, and R. Howe, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2014: 17th International Conference, Boston, MA, USA, September 14-18, 2014, Proceedings, Part II*, pages 349–356, Cham, 2014. Springer International Publishing.

[140] D. S. Marcus, A. F. Fotenos, J. G. Csernansky, J. C. Morris, and R. L. Buckner. Open access series of imaging studies: Longitudinal MRI data in nondemented and demented older adults. *Journal of Cognitive Neuroscience*, 22(12):2677–2684, 2010. PMID: 19929323.

[141] D. S. Marcus, T. H. Wang, J. Parker, J. G. Csernansky, J. C. Morris, and R. L. Buckner. Open access series of imaging studies (OASIS): Cross-sectional MRI data in young, middle aged, nondemented, and demented older adults. *Journal of Cognitive Neuroscience*, 19(9):1498–1507, 2007.

[142] R. Marsh, A. J. Gerber, and B. S. Peterson. Neuroimaging studies of normal brain development and their relevance for understanding childhood neuropsychiatric disorders. *Journal of the American Academy of Child & Adolescent Psychiatry*, 47(11):1233 – 1251, 2008.

[143] G. P. Mazzara, R. P. Velthuizen, J. L. Pearlman, H. M. Greenberg, and H. Wagner. Brain tumor target volume determination for radiation treatment planning through automated MRI segmentation. *International Journal of Radiation Oncology*Biology*Physics*, 59(1):300 – 312, 2004.

[144] J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, C. Holmes, L. Collins, P. Thompson, D. MacDonald, M. Iacoboni, T. Schormann, K. Amunts, N. Palomero-Gallagher, S. Geyer, L. Parsons, K. Narr, N. Kabani, G. L. Goualher, D. Boomsma, T. Cannon, R. Kawashima, and B. Mazoyer. A probabilistic atlas and reference system for the human brain: International consortium for brain mapping (ICBM). *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 356(1412):1293–1322, 2001.

[145] J. Mazziotta, A. Toga, A. Evans, P. Fox, J. Lancaster, K. Zilles, R. Woods, T. Paus, G. Simpson, B. Pike, C. Holmes, L. Collins, P. Thompson, D. MacDonald, M. Iacoboni, T. Schormann, K. Amunts, N. Palomero-Gallagher, S. Geyer, L. Parsons, K. Narr, N. Kabani, G. Le Goualher, J. Feidler, K. Smith, D. Boomsma, H. H. Pol, T. Cannon, R. Kawashima, and B. Mazoyer. A four-dimensional probabilistic atlas of the human brain. *Journal of the American Medical Informatics Association*, 8(5):401–430, 2001.

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

[147] B. C. Munsell, A. Temlyakov, and S. Wang. Fast multiple shape correspondence by pre-organizing shape instances. In *2009 IEEE Con-*

ference on Computer Vision and Pattern Recognition, pages 840–847, June 2009.

[148] V. Noblet, C. Heinrich, F. Heitz, and J.-P. Armspach. Symmetric nonrigid image registration: Application to average brain templates construction. In D. Metaxas, L. Axel, G. Fichtinger, and G. Székely, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008*, pages 897–904, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[149] N. Otsu. A threshold selection method from gray-level histograms. *IEEE transactions on systems, man, and cybernetics*, 9(1):62–66, 1979.

[150] Y. Ou, A. Sotiras, N. Paragios, and C. Davatzikos. DRAMMS: Deformable registration via attribute matching and mutual-saliency weighting. *Medical Image Analysis*, 15(4):622–639, 2010.

[151] A. Parraga, A. Susin, J. Pettersson, B. Macq, and M. D. Craene. 3D atlas building in the context of head and neck radiotherapy based on dense deformation fields. In *XX Brazilian Symposium on Computer Graphics and Image Processing (SIBGRAPI 2007)*, pages 321–328, Oct 2007.

[152] T. Paus, F. Tomaiuolo, N. Otaky, D. MacDonald, M. Petrides, J. Atlas, R. Morris, and A. C. Evans. Human cingulate and paracingulate sulci: pattern, variability, asymmetry, and probabilistic map. *Cereb Cortex*, 6(2):207–214, Mar-Apr 1996.

[153] X. Pennec, R. Stefanescu, V. Arsigny, P. Fillard, and N. Ayache. Riemannian elasticity: A statistical regularization framework for non-linear registration. In *Int. Conf. Med. Image Comput. Comput.-Assist. Intervent.*, pages 943–950. Springer, 2005.

[154] B. Peterson, M. A. Riddle, D. J. Cohen, L. D. Katz, J. C. Smith, M. T. Hardin, and J. F. Leckman. Reduced basal ganglia volumes in

Tourette's syndrome using threedimensional reconstruction techniques from magnetic resonance images. *Neurology*, 43(5):941, 1993.

[155] B. S. Peterson, H. A. Choi, X. Hao, J. A. Amat, H. Zhu, R. Whiteman, J. Liu, D. Xu, and R. Bansal. Morphologic features of the amygdala and hippocampus in children and adults with Tourette syndrome. *Archives of General Psychiatry*, 64(11):1281–1291, 2007.

[156] B. S. Peterson, P. Thomas, M. J. Kane, L. Scahill, H. Zhang, R. Bronen, R. A. King, J. F. Leckman, and L. Staib. Basal ganglia volumes in patients with Gilles de la Tourette syndrome. *Archives of General Psychiatry*, 60(4):415–424, 2003.

[157] C. Platero and M. C. Tobar. A fast approach for hippocampal segmentation from T1-MRI for predicting progression in Alzheimer's disease from elderly controls. *Journal of Neuroscience Methods*, 270:61 – 75, 2016.

[158] J. M. Prewitt. Object enhancement and extraction. *Picture processing and Psychopictorics*, 10(1):15–19, 1970.

[159] R. D. Rabbitt, J. A. Weiss, G. E. Christensen, and M. I. Miller. Mapping of hyperelastic deformable templates using the finite element method. *Proc.SPIE*, 2573:2573 – 2573 – 14, 1995.

[160] J. C. Rajapakse, J. N. Giedd, and J. L. Rapoport. Statistical approach to segmentation of single-channel cerebral MR images. *IEEE Transactions on Medical Imaging*, 16(2):176–186, April 1997.

[161] L. Ramus, O. Commowick, and G. Malandain. Construction of patient specific atlases from locally most similar anatomical pieces. In T. Jiang, N. Navab, J. P. W. Pluim, and M. A. Viergever, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2010: 13th International Conference, Beijing, China, September 20-24, 2010,*

Proceedings, Part III, pages 155–162, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[162] L. Ramus and G. Malandain. Assessing selection methods in the context of multi-atlas based segmentation. In *2010 IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 1321–1324, April 2010.

[163] L. Ramus and G. Malandain. Multi-atlas based segmentation: Application to the head and neck region for radiotherapy planning. In *MICCAI Workshop Medical Image Analysis for the Clinic-A Grand Challenge*, pages 281–288, 2010.

[164] T. Ridler and S. Calvard. Picture thresholding using an iterative selection method, 08 1978.

[165] V. Roessner, S. Overlack, C. Schmidt-Samoa, J. Baudewig, P. Dechent, A. Rothenberger, and G. Helms. Increased putamen and callosal motor subregion in treatment-naive boys with Tourette syndrome indicates changes in the bihemispheric motor network. *Journal of Child Psychology and Psychiatry*, 52(3):306–314, 2011.

[166] J. Rogowska. Chapter 5 - Overview and fundamentals of medical image segmentation. In I. N. BANKMAN, editor, *Handbook of Medical Image Processing and Analysis (Second Edition)*, pages 73 – 90. Academic Press, Burlington, second edition edition, 2009.

[167] T. Rohlfing. Image similarity and tissue overlaps as surrogates for image registration accuracy: Widely used but unreliable. *Medical Imaging, IEEE Transactions on*, 31(2):153–163, Feb 2012.

[168] T. Rohlfing, R. Brandt, R. Menzel, and C. Maurer. Segmentation of three-dimensional images using non-rigid registration: Methods and validation with application to confocal microscopy images of bee brains.

In *IN MEDICAL IMAGING: IMAGE PROCESSING*, pages 363–374, 2003.

[169] T. Rohlfing, R. Brandt, R. Menzel, and C. R. Maurer. Evaluation of atlas selection strategies for atlas-based image segmentation with application to confocal microscopy images of bee brains. *NeuroImage*, 21(4):1428 – 1442, 2004.

[170] T. Rohlfing, R. Brandt, R. Menzel, D. B. Russakoff, and C. R. Maurer. Quo vadis, atlas-based segmentation? In J. S. Suri, D. L. Wilson, and S. Laxminarayan, editors, *Handbook of Biomedical Image Analysis: Volume III: Registration Models*, pages 435–486, Boston, MA, 2005. Springer US.

[171] T. Rohlfing and C. R. Maurer. Multi-classifier framework for atlas-based image segmentation. *Pattern Recognition Letters*, 26(13):2070 – 2079, 2005.

[172] T. Rohlfing, D. B. Russakoff, and C. R. Maurer. An expectation maximization-like algorithm for multi-atlas multi-label segmentation. In T. Wittenberg, P. Hastreiter, U. Hoppe, H. Handels, A. Horsch, and H.-P. Meinzer, editors, *Bildverarbeitung für die Medizin 2003: Algorithmen — Systeme — Anwendungen, Proceedings des Workshops vom 9.–11. März 2003 in Erlangen*, pages 348–352, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[173] T. Rohlfing, D. B. Russakoff, and C. R. Maurer. Extraction and application of expert priors to combine multiple segmentations of human brain tissue. In R. E. Ellis and T. M. Peters, editors, *Medical Image Computing and Computer-Assisted Intervention - MICCAI 2003: 6th International Conference, Montréal, Canada, November 15-18, 2003. Proceedings*, pages 578–585, Berlin, Heidelberg, 2003. Springer Berlin Heidelberg.

[174] F. Rousseau, P. Habas, and C. Studholme. A supervised patch-based approach for human brain labeling. *Medical Imaging, IEEE Transactions on*, 30(10):1852–1862, Oct 2011.

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

[176] D. Rueckert, P. Aljabar, R. A. Heckemann, J. V. Hajnal, and A. Hammers. Diffeomorphic registration using B-splines. In R. Larsen, M. Nielsen, and J. Sporring, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2006*, pages 702–709, Berlin, Heidelberg, 2006. Springer Berlin Heidelberg.

[177] 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.

[178] M. R. Sabuncu, B. T. T. Yeo, K. V. 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, Oct 2010.

[179] S. Sandor and R. Leahy. Surface-based labeling of cortical anatomy using a deformable atlas. *IEEE Transactions on Medical Imaging*, 16(1):41–54, Feb 1997.

[180] G. Sanroma, G. Wu, Y. Gao, and D. Shen. Learning to rank atlases for multiple-atlas segmentation. *Medical Imaging, IEEE Transactions on*, 33(10):1939–1953, Oct 2014.

[181] G. Sanroma, G. Wu, K. Thung, Y. Guo, and D. Shen. Novel multi-atlas segmentation by matrix completion. In G. Wu, D. Zhang, and L. Zhou, editors, *Machine Learning in Medical Imaging*, volume 8679 of *LNCS*, pages 207–214. Springer International Publishing, 2014.

[182] J. M. Scharf, L. L. Miller, C. A. Mathews, and Y. Ben-Shlomo. Prevalence of Tourette syndrome and chronic tics in the population-based avon longitudinal study of parents and children cohort. *Journal of the American Academy of Child & Adolescent Psychiatry*, 51(2):192–201.e5, 09 2011.

[183] M. Sdika. Combining atlas based segmentation and intensity classification with nearest neighbor transform and accuracy weighted vote. *Medical Image Analysis*, 14(2):219 – 226, 2010.

[184] A. Serag, P. Aljabar, G. Ball, S. J. Counsell, J. P. Boardman, M. A. Rutherford, A. D. Edwards, J. V. Hajnal, and D. Rueckert. Construction of a consistent high-definition spatio-temporal atlas of the developing brain using adaptive kernel regression. *NeuroImage*, 59(3):2255 – 2265, 2012.

[185] D. W. Shattuck, M. Mirza, V. Adisetiyo, C. Hojatkashani, G. Salamon, K. L. Narr, R. A. Poldrack, R. M. Bilder, and A. W. Toga. Construction of a 3D probabilistic atlas of human cortical structures. *NeuroImage*, 39(3):1064 – 1080, 2008.

[186] D. W. Shattuck, S. R. Sandor-Leahy, K. A. Schaper, D. A. Rottenberg, and R. M. Leahy. Magnetic resonance image tissue classification using a partial volume model. *NeuroImage*, 13(5):856 – 876, 2001.

[187] F. Shi, P.-T. Yap, Y. Fan, J. H. Gilmore, W. Lin, and D. Shen. Construction of multi-region-multi-reference atlases for neonatal brain MRI segmentation. *NeuroImage*, 51(2):684 – 693, 2010.

[188] C. Sjöberg, S. Johansson, and A. Ahnesjö. How much will linked deformable registrations decrease the quality of multi-atlas segmentation fusions? *Radiation Oncology*, 9(1):251, Dec 2014.

[189] C. Sjberg and A. Ahnesj. Multi-atlas based segmentation using probabilistic label fusion with adaptive weighting of image similarity mea-

sures. *Computer Methods and Programs in Biomedicine*, 110(3):308 – 319, 2013.

[190] 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.

[191] S. M. Smith. Fast robust automated brain extraction. *Human Brain Mapping*, 17(3):143–155, 2002.

[192] I. Sobel. An isotropic 3 3 image gradient operator, 02 1968.

[193] H. Sokooti, G. Saygili, B. Glocker, B. P. F. Lelieveldt, and M. Staring. Accuracy estimation for medical image registration using regression forests. In S. Ourselin, L. Joskowicz, M. R. Sabuncu, G. Unal, and W. Wells, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2016: 19th International Conference, Athens, Greece, October 17-21, 2016, Proceedings, Part III*, pages 107–115, Cham, 2016. Springer International Publishing.

[194] G. Song, B. B. Avants, and J. C. Gee. Multi-start method with prior learning for image registration. In *2007 IEEE 11th International Conference on Computer Vision*, pages 1–8, Oct 2007.

[195] J. H. Song, G. E. Christensen, J. A. Hawley, Y. Wei, and J. G. Kuhl. Evaluating image registration using NIREP. In B. Fischer, B. M. Dawant, and C. Lorenz, editors, *Biomedical Image Registration: 4th International Workshop, WBIR 2010, Lübeck, Germany, July 11-13, 2010. Proceedings*, pages 140–150, Berlin, Heidelberg, 2010. Springer Berlin Heidelberg.

[196] A. Sotiras, C. Davatzikos, and N. Paragios. Deformable medical image registration: A survey. *IEEE transactions on medical imaging*, 32(7):1153–1190, 2013.

[197] E. R. Sowell, E. Kan, J. Yoshii, P. M. Thompson, R. Bansal, D. Xu, A. W. Toga, and B. S. Peterson. Thinning of sensorimotor cortices in children with Tourette syndrome. *Nat Neurosci*, 11(6):637–639, 06 2008.

[198] E. Stern, D. Silbersweig, K. Chee, A. Holmes, M. M. Robertson, M. Trimble, C. D. Frith, S. J. Frackowiak, and R. J. Dolan. A functional neuroanatomy of tics in Tourette syndrome. *Archives of General Psychiatry*, 57(8):741–748, 2000.

[199] L. Sun, C. Zu, and D. Zhang. Reliability guided forward and backward patch-based method for multi-atlas segmentation. In G. Wu, P. Coupé, Y. Zhan, B. Munsell, and D. Rueckert, editors, *Patch-Based Techniques in Medical Imaging: First International Workshop, Patch-MI 2015, Held in Conjunction with MICCAI 2015, Munich, Germany, October 9, 2015, Revised Selected Papers*, pages 128–136, Cham, 2015. Springer International Publishing.

[200] F. Sgonne, A. Dale, E. Busa, M. Glessner, D. Salat, H. Hahn, and B. Fischl. A hybrid approach to the skull stripping problem in MRI. *NeuroImage*, 22(3):1060 – 1075, 2004.

[201] J. Talairach, M. David, P. Tournoux, H. Corredor, and T. Kvasina. Atlas d'anatomie stereotaxique des noyaux gris centraux, 1957.

[202] S. Tang, Y. Fan, G. Wu, M. Kim, and D. Shen. Rabbit: Rapid alignment of brains by building intermediate templates. *NeuroImage*, 47(4):1277 – 1287, 2009.

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

[204] C.-C. Teng, L. G. Shapiro, and I. J. Kalet. Head and neck lymph node

region delineation with image registration. *BioMedical Engineering OnLine*, 9(1):30, Jun 2010.

[205] J.-P. Thirion. Image matching as a diffusion process: an analogy with Maxwell's demons. *Medical Image Analysis*, 2(3):243 – 260, 1998.

[206] P. Thompson, K. Hayashi, E. Sowell, N. Gogtay, J. Giedd, J. Rapoport, G. De Zubicaray, A. Janke, S. Rose, J. Semple, D. Doddrell, Y. Wang, T. Van Erp, T. Cannon, and A. Toga. Mapping cortical change in Alzheimer's disease, brain development, and schizophrenia. *NeuroImage*, 23(SUPPL. 1), 2004.

[207] P. Thompson, C. Schwartz, and A. Toga. High-resolution random mesh algorithms for creating a probabilistic 3D surface atlas of the human brain. *NeuroImage*, 3(1):19 – 34, 1996.

[208] P. M. Thompson, D. MacDonald, M. S. Mega, C. J. Holmes, A. C. Evans, and A. W. Toga. Detection and mapping of abnormal brain structure with a probabilistic atlas of cortical surfaces. *Journal of computer assisted tomography*, 21(4):567–581, 1997.

[209] P. M. Thompson and A. W. Toga. Detection, visualization and animation of abnormal anatomic structure with a deformable probabilistic brain atlas based on random vector field transformations. *Medical Image Analysis*, 1(4):271 – 294, 1997.

[210] K. D. Toennies. *Segmentation: Principles and Basic Techniques*, pages 171–209. Springer London, London, 2012.

[211] A. Trouvé. Diffeomorphisms groups and pattern matching in image analysis. *International Journal of Computer Vision*, 28(3):213–221, Jul 1998.

[212] K. P. Tung, W. J. Bei, W. Z. Shi, H. Y. Wang, T. Tong, R. D. Silva, E. Edwards, and D. Rueckert. Multi-atlas based neointima segmenta-

tion in intravascular coronary oct. In *2013 IEEE 10th International Symposium on Biomedical Imaging*, pages 1280–1283, April 2013.

[213] M. Turk and A. Pentland. Eigenfaces for recognition. *Journal of Cognitive Neuroscience*, 3(1):71–86, 1991. PMID: 23964806.

[214] N. J. Tustison, B. B. Avants, P. A. Cook, Y. Zheng, A. Egan, P. A. Yushkevich, and J. C. Gee. N4ITK: Improved N3 bias correction. *IEEE Transactions on Medical Imaging*, 29(6):1310–1320, June 2010.

[215] S. Valverde, A. Oliver, M. Cabezas, E. Roura, and X. Llad. Comparison of 10 brain tissue segmentation methods using revisited IBSR annotations. *Journal of Magnetic Resonance Imaging*, 41(1):93–101, 2015.

[216] 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.

[217] E. M. van Rikxoort, I. Isgum, Y. Arzhaeva, M. Staring, S. Klein, M. A. Viergever, J. P. Pluim, and B. van Ginneken. Adaptive local multi-atlas segmentation: Application to the heart and the caudate nucleus. *Medical Image Analysis*, 14(1):39 – 49, 2010.

[218] T. Vercauteren, X. Pennec, A. Perchant, and N. Ayache. Symmetric log-domain diffeomorphic registration: A demons-based approach. In D. Metaxas, L. Axel, G. Fichtinger, and G. Székely, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2008*, pages 754–761, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg.

[219] J. Wan, A. Carass, S. M. Resnick, and J. L. Prince. Automated reliable labeling of the cortical surface. In *2008 5th IEEE International Symposium on Biomedical Imaging: From Nano to Macro*, pages 440–443, May 2008.

[220] H. Wang, S. R. Das, J. W. Suh, M. Altinay, J. Pluta, C. Craige, B. Avants, and P. A. Yushkevich. A learning-based wrapper method to correct systematic errors in automatic image segmentation: Consistently improved performance in hippocampus, cortex and brain segmentation. *NeuroImage*, 55(3):968 – 985, 2011.

[221] H. Wang, J. W. Suh, S. R. Das, J. B. Pluta, C. Craige, and P. A. Yushkevich. Multi-atlas segmentation with joint label fusion. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(3):611–623, March 2013.

[222] H. Wang and P. Yushkevich. Multi-atlas segmentation with joint label fusion and corrective learning - an open source implementation. *Frontiers in Neuroinformatics*, 7:27, 2013.

[223] H. Wang and P. A. Yushkevich. Groupwise segmentation with multi-atlas joint label fusion. In K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part I*, pages 711–718. Springer Berlin Heidelberg, Berlin, Heidelberg, 2013.

[224] H. Wang and P. A. Yushkevich. Multi-atlas segmentation without registration: A supervoxel-based approach. In K. Mori, I. Sakuma, Y. Sato, C. Barillot, and N. Navab, editors, *Medical Image Computing and Computer-Assisted Intervention – MICCAI 2013: 16th International Conference, Nagoya, Japan, September 22-26, 2013, Proceedings, Part III*, pages 535–542, Berlin, Heidelberg, 2013. Springer Berlin Heidelberg.

[225] Q. Wang, L. Chen, P.-T. Yap, G. Wu, and D. Shen. Groupwise registration based on hierarchical image clustering and atlas synthesis. *Human Brain Mapping*, 31(8):1128–1140, 2010.

[226] Q. Wang, G. Wu, M.-J. Kim, L. Zhang, and D. Shen. Interactive registration and segmentation for multi-atlas-based labeling of brain MR image. In H. Zha, X. Chen, L. Wang, and Q. Miao, editors, *Computer Vision: CCF Chinese Conference, CCCV 2015, Xi'an, China, September 18-20, 2015, Proceedings, Part I*, pages 240–248, Berlin, Heidelberg, 2015. Springer Berlin Heidelberg.

[227] Z. Wang, A. C. Bovik, H. R. Sheikh, and E. P. Simoncelli. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612, April 2004.

[228] S. K. Warfield, K. H. Zou, and W. M. Wells. Validation of image segmentation and expert quality with an expectation-maximization algorithm. In T. Dohi and R. Kikinis, editors, *Medical Image Computing and Computer-Assisted Intervention — MICCAI 2002: 5th International Conference Tokyo, Japan, September 25–28, 2002 Proceedings, Part I*, pages 298–306, Berlin, Heidelberg, 2002. Springer Berlin Heidelberg.

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

[230] W. M. Wells, P. Viola, H. Atsumi, S. Nakajima, and R. Kikinis. Multimodal volume registration by maximization of mutual information. *Medical image analysis*, 1(1):35–51, 1996.

[231] C. Westbury, R. Zatorre, and A. Evans. Quantifying variability in the planum temporale: A probability map. *Cerebral Cortex*, 9(4):392–405, 1999.

[232] D. Withey and Z. Koles. Medical image segmentation: Methods and software. In *Noninvasive Functional Source Imaging of the Brain*

and Heart and the International Conference on Functional Biomedical Imaging, 2007. NFSI-ICFBI 2007. Joint Meeting of the 6th International Symposium on, pages 140–143. IEEE, 2007.

[233] M. Wittfoth, S. Bornmann, T. Peschel, J. Grosskreutz, A. Glahn, N. Buddensiek, H. Becker, R. Dengler, and K. R. Müller-Vahl. Lateral frontal cortex volume reduction in Tourette syndrome revealed by VBM. *BMC Neuroscience*, 13(1):17, Feb 2012.

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

[235] R. Wolz, C. Chu, K. Misawa, M. Fujiwara, K. Mori, and D. Rueckert. Automated abdominal multi-organ segmentation with subject-specific atlas generation. *IEEE Transactions on Medical Imaging*, 32(9):1723–1730, Sept 2013.

[236] G. Wu, M. Kim, G. Sanroma, Q. Wang, B. C. Munsell, and D. Shen. Hierarchical multi-atlas label fusion with multi-scale feature representation and label-specific patch partition. *NeuroImage*, 106:34 – 46, 2015.

[237] G. Wu, Q. Wang, Daoqiang, and D. Shen. Robust patch-based multi-atlas labeling by joint sparsity regularization. In *MICCAI 2012 Workshop STMI*, pages 91–94, 2012.

[238] G. Wu, Q. Wang, D. Zhang, F. Nie, H. Huang, and D. Shen. A generative probability model of joint label fusion for multi-atlas based brain segmentation. *Medical Image Analysis*, 18(6):881 – 890, 2014.

[239] 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.

[240] Q. Xie and D. Ruan. Low-complexity atlas-based prostate segmentation by combining global, regional, and local metrics. *Medical Physics*, 41(4):041909–n/a, 2014. 041909.

[241] H. Xue, L. Srinivasan, S. Jiang, M. Rutherford, A. D. Edwards, D. Rueckert, and J. V. Hajnal. Automatic segmentation and reconstruction of the cortex from neonatal MRI. *NeuroImage*, 38(3):461 – 477, 2007.

[242] P.-Y. Yin and L.-H. Chen. A fast iterative scheme for multilevel thresholding methods. *Signal Processing*, 60(3):305 – 313, 1997.

[243] D. Zhang, Q. Guo, G. Wu, and D. Shen. Sparse patch-based label fusion for multi-atlas segmentation. In P.-T. Yap, T. Liu, D. Shen, C.-F. Westin, and L. Shen, editors, *Multimodal Brain Image Analysis: Second International Workshop, MBIA 2012, Held in Conjunction with MICCAI 2012, Nice, France, October 1-5, 2012. Proceedings*, pages 94–102, Berlin, Heidelberg, 2012. Springer Berlin Heidelberg.

[244] Y. Zhang, M. Brady, and S. Smith. Segmentation of brain MR images through a hidden Markov random field model and the expectation-maximization algorithm. *IEEE Transactions on Medical Imaging*, 20(1):45–57, Jan 2001.

[245] D. Zikic, B. Glocker, and A. Criminisi. Encoding atlases by randomized classification forests for efficient multi-atlas label propagation. *Medical Image Analysis*, 18(8):1262 – 1273, 2014. Special Issue on the MICCAI 2013.